A Deep Latent Factor Graph Clustering with Fairness-Utility Trade-off Perspective

Siamak Ghodsi [ID]
L3S Research Center
Hannover, Germany
ghodsi[at]l3s.de

Amjad Seyedi [ID]
University of Mons
Mons, Belgium
seyedamjad.seyedi[at]umons.ac.be

Tai Le Quy [ID]
University of Koblenz
Koblenz, Germany
tailequy[at]uni-koblenz.de

Fariba Karimi [ID]
TU Graz
Graz, Austria
karimi[at]csh.ac.at

Eirini Ntoutsi [ID]
Bundeswehr University
Munich, Germany
eirini.ntoutsi[at]unibw.de

Abstract—**Fair graph clustering seeks partitions that respect network structure while maintaining proportional representation across sensitive groups, with applications spanning community detection, team formation, resource allocation, and social network analysis. Many existing approaches enforce rigid constraints or rely on multi-stage pipelines (e.g., spectral embedding followed by k-means), limiting trade-off control, interpretability, and scalability. We introduce *DFNMF*, an end-to-end deep nonnegative tri-factorization tailored to graphs that directly optimizes cluster assignments with a soft statistical-parity regularizer. A single parameter λ tunes the fairness–utility balance, while nonnegativity yields parts-based factors and transparent soft memberships. The optimization uses sparse-friendly alternating updates and scales near-linearly with the number of edges. Across synthetic and real networks, DFNMF achieves substantially higher group balance at comparable modularity, often dominating state-of-the-art baselines on the Pareto front. The code is available at https://github.com/SiamakGhodsi/DFNMF.git.**
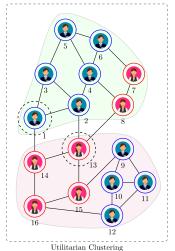
Index Terms—**Trustworthy ML, Fair Graph Clustering, Community Detection, Heterogen. Networks, Deep Factorization**

I. INTRODUCTION

Fair and trustworthy machine learning has advanced rapidly over the last decade [1]–[4], yet it remains underexplored in graph learning—especially graph clustering [5], [6]. Graph clustering partitions a network into cohesive groups and underpins applications such as community detection, team formation, resource allocation, and social network analysis [7]. Classical approaches maximize utility (e.g., modularity) or minimize cuts to uncover natural network structures, yet many real-world applications require balancing structural cohesion with demographic fairness.

Consider academic collaboration networks where funders increasingly mandate diverse team compositions. Purely structure-driven clustering may recover prolific yet homogeneous groups that fail diversity requirements. Similarly, educational institutions forming student project teams must balance social connections with equitable assignments (Figure 1). These scenarios demand fair graph clustering—modifying natural community boundaries to achieve demographic balance while preserving meaningful network structure [8].

Two lines of research relate to our work. **(i) Fair clustering for i.i.d. data**: develop demographic-parity constraints for k-center/median/means via fairlets and related schemes [9], [10], alongside supervised fair GNNs for labeled settings [11]–[13]. These methods assume metric i.i.d. data or require labels,
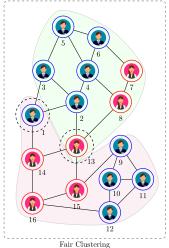

Fig. 1: Fair clustering of a 16-node graph (10 **M**ale, 6 **F**emale) into two equal-sized clusters. Left: Utilitarian clustering yields a structure-driven partition with a 6M:2F distribution for green and 4M:4F for lavender cluster, resulting in gender imbalance. Right: Fair clustering achieves a balanced 5M:3F distribution in both clusters by swapping memberships of nodes 1 and 13.

making them unsuitable for graph-structured clustering. **(ii) Fair graph clustering**: primarily extends spectral clustering with hard demographic constraints [14]–[16], but eigen-decompositions and post-hoc rounding (e.g., k-means) limit end-to-end control and scalability. Recent matrix-factorization approaches offer more flexibility, though contrastive NMF variants can struggle at larger scales [17].

To address these limitations, we propose **DFNMF**, an end-to-end deep NMF framework for fair graph clustering. DFNMF integrates balanced fairness constraints—enforcing proportional group representation—directly into the clustering objective, enabling explicit utility–fairness trade-offs via a single parameter λ without post-processing. Deep tri-factorization captures multi-level network structure while maintaining scalability through sparse implementations. To our knowledge, this is the first deep hierarchical NMF model with integrated fairness designed specifically for graph clustering. Our contributions are:

- **End-to-end fairness integration:** Joint representation learning and fairness enforcement without multi-stage

pipelines or post-processing.

- **Tunable trade-off control:** Single parameter λ enables precise utility–fairness balance adjustment.
- **Deep hierarchical architecture:** Tri-factorization robust to graph sparsity and group-size heterogeneity.
- **Scalable sparse implementation:** CSR-based operations with efficient alternating updates.
- **Comprehensive evaluation:** Synthetic and real networks with statistical analysis, Pareto studies, and ablations.

The remainder of this paper is structured as follows: Section II reviews the current literature and highlights the existing gaps. Section III formulates the problem and provides necessary preliminaries, including theoretical foundations of NMF. Section IV details our proposed DFNMF model. Section V presents the experimental results. Finally, Section VI concludes the paper and points out future research directions.

II. RELATED WORKS

Fairness-aware graph clustering aims to mitigate bias propagation by enforcing demographic constraints during network partitioning. While graph fairness has been studied extensively [5], [7], [18], relatively few works specifically tackle fairness in unsupervised graph clustering. The literature can be broadly divided into (1) methods primarily targeting independent and identically distributed (iid) data, and (2) those tailored specifically for graph-structured data.

A. Fair Clustering for iid Data

Fairness-aware clustering methods initially emerged for independent and identically distributed (iid) tabular data. Chierichetti et al. [9] introduced "fairlets" to enforce demographic parity in k-center and k-median clustering. Subsequent works extended fairness constraints to k-means and related objectives [10]. Some approaches construct similarity graphs from iid data to enable graph-based fairness reasoning [19], [20]. However, these methods rely on strong geometric or metric assumptions and cannot natively handle the complex relational dependencies inherent in real-world networks [5], [21].

While these methods address unsupervised fairness, they are fundamentally limited for graph clustering applications. Their experimental evaluations typically use simplified datasets that inadequately reflect the structural complexities of networked data, where connections themselves carry semantic meaning beyond mere similarity.

B. Fair Graph Clustering

Recent studies directly extend spectral clustering to incorporate fairness constraints for graph data. Kleindessner et al. [14] introduced demographic fairness constraints into spectral decompositions. Subsequent works improved computational efficiency using relaxation techniques and augmented Lagrangian methods [15], [16], [22]. However, these spectral methods face fundamental limitations: rigid fairness constraints, non-end-to-end clustering requiring post-processing (e.g., k-means), discretization approximations, and limited interpretability.

Graph embedding-based alternatives such as the method by Dong et al. [11] apply individual fairness through rank-based alignment, yet their embedding-driven strategy risks propagating biases from sensitive attributes. Ghodsi et al. [17] proposed asymmetric NMF with contrastive fairness regularization, offering improved interpretability but limited scalability due to computationally intensive pairwise contrastive terms.

Supervised graph methods have also been explored for fairness-aware tasks. Ranking-based fairness models [11], fair GNN architectures [12], and sensitive attribute neutralization strategies [13] show promise in supervised settings. However, these approaches require labeled data and may suffer from proxy-bias propagation through learned embeddings, limiting their applicability to unsupervised community discovery tasks.

GNN-based clustering methods like DMoN [23] provide scalable alternatives but lack integrated fairness constraints. While effective for supervised tasks, adapting such methods to incorporate fairness remains an open challenge.

Table I summarizes existing methods across five key criteria. Most approaches either face scalability limitations, enforce rigid fairness constraints, or provide limited interpretability. Existing approaches trade off flexibility (softness), end-to-end optimization, and scalability—gaps our DFNMF addresses.

Clearly, existing methods either face scalability constraints, rigid fairness enforcement, or limited interpretability.

Method	Soft?	End-to-End	Complexity	Scalability	Interpret.
[23] DMoN	✔	✔	$O(nk^2 + |E|)$	Large ↑	Low ↓
[14] FSC	✗	✗	$O(n^3)$	Small ↓	Low ↓
[16] sFSC	✗	✗	$O(|E| + n(h^2 + k^2))$	Medium	Low↓
[15] iFSC	✗	✗	$O(n^3)$	Small ↓	Low ↓
[11] GNN-FSC	✗	✗	$O(n^3)$	Small ↓	Low ↓
[22] FNM-SC	✗	✗	$O(n^{4.5}k^{4.5}|E|)$	Small ↓	Low ↓
[17] iFNMTF	✔	✔	$O(|E|k + nk^2 + k^3)$	Medium	High ↑
DFNMF (ours)	✔	✔	$O(Tpk|E|)$	Large ↑	High ↑

TABLE I: Representative methods for fair graph clustering. *Soft?* indicates soft (regularized) vs. hard constraints; *End-to-End* denotes no post-hoc rounding; complexity is measured in graph size n, edges $|E|$, clusters k, and iterations T, and *Interpret.* refers to parts-based transparency of assignments/factors.

III. PROBLEM FORMULATION & PRELIMINARIES

A. Problem Definition

Consider an undirected graph $\mathcal{G} = (V, E)$ with adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ encoding edge connections, where $a_{ij} > 0$ indicates a positive edge between nodes $v_i, v_j \in V$, and $a_{ij} = 0$ otherwise (assuming no self-loops, thus $a_{ii} = 0$). Suppose the node set V is partitioned into m disjoint demographic groups based on sensitive attributes, such that $V = \dot{\cup}_{s \in [m]} V_s$. Our goal is to find a clustering $C = C_1 \dot{\cup} \ldots \dot{\cup} C_k$ into k clusters that maximizes structural cohesion while satisfying demographic balance (Definition 1), ensuring fair representation.

B. Demographic Balance

The *balance* criterion, originally introduced by Chierichetti et al. [9], ensures that each cluster maintains demographic

proportions similar to the global distribution. This concept, first applied to k-median, was later adapted to spectral clustering in [14], [16]. While initially formulated for iid data, we extend this principle to our context, where demographic balance must be preserved over node partitions, formalized as follows:

Definition 1 (Generalized Demographic Balance)**.** *Given a partitioning of the vertex set V into $k \geq 2$ communities, the clustering is said to be* fair *with respect to a partition into demographic groups $s \in [m]$ if, in each community C_l, the proportion of group-s nodes matches its global share:*

$$\frac{|V_s \cap C_l|}{|C_l|} = \frac{|V_s|}{|V|}, \quad \forall s \in \{1, \ldots, m\}, \ l \in \{1, \ldots, k\}. \quad (1)$$

Here, $V = \{v_1, v_2, \ldots, v_n\}$ is the set of graph nodes, $|V_s|$ is the number of nodes belonging to demographic group s, and $|V| = n$ is the total number of nodes. This constraint operationalizes *statistical parity* for clustering and forms the foundation of the fairness regularizer introduced in Section IV.

C. *Nonnegative Matrix Factorization (NMF)*

NMF is widely used due to its interpretability, versatility, and adaptability to constraints, performing a low-rank approximation of data [24]. Its nonnegativity constraints yield parts-based, sparse representations that are inherently interpretable, making NMF particularly suitable for domains like text mining, image analysis, and bioinformatics [25], [26]. Formally, given data matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$, NMF seeks basis matrix $\boldsymbol{W} \in \mathbb{R}^{m \times k}$ and representation matrix $\boldsymbol{H} \in \mathbb{R}^{k \times n}$ by solving:

$$\min_{\boldsymbol{W}, \boldsymbol{H} \geq 0} \|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F^2, \quad (2)$$

where $\|\boldsymbol{A}\|_F = \sqrt{\sum_{ij} A_{ij}^2}$ denotes the Frobenius norm.

1) NMF Tri-Factorization for Graph Clustering: The general NMF formulation (2) is not directly tailored for graph clustering. Hence, Symmetric-NMF [27] was developed specifically for this purpose. Symmetric-NMF factorizes a graph adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ into latent node representations $\boldsymbol{H} \in \mathbb{R}^{n \times u}$, encouraging attraction between connected nodes and repulsion otherwise. Formally:

$$\min_{\boldsymbol{H} \geq 0} \|\boldsymbol{A} - \boldsymbol{H}\boldsymbol{H}^\top\|_F^2. \quad (3)$$

Extending Symmetric NMF, the NMF Tri-Factorization (NMTF) [28] introduces cluster-cluster interactions via matrix $\boldsymbol{W}$, leading to:

$$\min_{\boldsymbol{H}, \boldsymbol{W} \geq 0} \|\boldsymbol{A} - \boldsymbol{H}\boldsymbol{W}\boldsymbol{H}^\top\|_F^2, \quad (4)$$

where $\boldsymbol{W}$ represents interactions between clusters, improving interpretability and flexibility.

An epistemic comparison of theoretical foundations and differences of NMF-based clustering vs neural network architectures is provided in the Appendix VIII.

IV. DEEP FAIR NMF MODEL

This section presents **DFNMF**, a deep fair tri-factor model for community detection on attributed graphs. A proportional group-balance regularizer is built into the objective, producing soft memberships and *direct* cluster assignments—no post-hoc k-means. The design (i) performs end-to-end clustering with interpretable nonnegative factors; (ii) couples graph topology with demographic indicators to promote balanced communities; (iii) traverses the utility–fairness trade-off via a single parameter λ; and (iv) scales through sparse-friendly alternating updates with shallow pretraining and deep fine-tuning.

A. *Soft Balanced Fairness (BF)*

We encode node–cluster memberships by $\boldsymbol{H} \in \mathbb{R}^{n \times k}$. For partitioning V into k clusters, projecting $\boldsymbol{H}$ by node memberships yields:

$$\sum_{i=1}^{n_l} H_{il} = 1, \quad \forall l \in \{1, \ldots, k\}. \quad (5)$$

This encoding enforces nonnegativity, improving interpretability via parts-based memberships. With column-stochastic normalization (Eq. (5)), each entry $H_{il} \in [0, 1]$ serves as a soft membership weight for node v_i in cluster C_l.

Definition 2 (Demographic Group Encoding)**.** *Let the vertex set V be partitioned into m sensitive groups $V = \dot{\cup}_{s=1}^m V_s$, and define the binary sensitive group indicator for node v_i as*

$$g_i^{(s)} = \begin{cases} 1, & \text{if } v_i \in V_s, \\ 0, & \text{otherwise.} \end{cases}$$

We then define the group indicator for each sensitive group as

$$\boldsymbol{f}^{(s)} = \boldsymbol{g}^{(s)} - \frac{|V_s|}{n} \boldsymbol{1}_n, \quad s \in \{1, \ldots, m-1\} \quad (6)$$

where $\boldsymbol{1}_n \in \mathbb{R}^n$ is the vector of ones. We obtain the matrix $\boldsymbol{F} = \begin{bmatrix} \boldsymbol{f}^{(1)} & \boldsymbol{f}^{(2)} & \cdots & \boldsymbol{f}^{(m-1)} \end{bmatrix} \in \mathbb{R}^{n \times (m-1)}$ by stacking these $m - 1$ vectors as columns. Note: We use $m - 1$ columns to avoid linear dependency, since group proportions sum to one.

The matrix $\boldsymbol{F}$ stacks mean-centered (proportionally centered) group indicators. It will encode demographic balance via the linear condition $\boldsymbol{F}^\top \boldsymbol{H} = \boldsymbol{0}$ and induce the fairness regularizer $\|\boldsymbol{F}^\top \boldsymbol{H}\|_F^2$. This construction is used as Step 1 in Algorithm 1.

Definition 3 (Soft Balanced Fairness)**.** *A partitioning represented by a column-stochastic membership matrix $\boldsymbol{H}$ is* fair *with respect to the sensitive groups if, for each cluster C_l and every group $s \in \{1, \ldots, m\}$, the proportion of mass from group V_s in C_l equals the global group proportion. It means that, a clustering $\boldsymbol{H}$ is fair if eq.(1) strictly holds for all clusters and all sensitive groups. In our soft clustering formulation, this condition is equivalent to the following:*

$$\sum_{i \in V_s} H_{il} = \frac{|V_s|}{n} \quad \text{for all } s \in \{1, \ldots, m-1\}, \ l \in \{1, \ldots, k\}.$$

Equivalently, stacking the deviations for all groups and clusters, the fairness condition can be formulated as $\boldsymbol{F}^\top \boldsymbol{H} = \boldsymbol{0}$.

Lemma 1 (Equivalence to Demographic Balance). *Let $\boldsymbol{H}$ be a nonnegative, column-stochastic matrix. Then $\boldsymbol{F}^\top \boldsymbol{H} = \boldsymbol{0}$ of Definition 3, is equivalent to the fairness condition in Eq.* (1).

Proof. Since each column of $\boldsymbol{H}$ is normalized by Eq. (5), for any cluster l we have $\sum_{i=1}^n H_{il} = 1$. Then, for each sensitive group $s \in [m-1]$ and cluster l, the (s,l)th entry of $\boldsymbol{F}^\top \boldsymbol{H}$ is

$$(\boldsymbol{F}^\top \boldsymbol{H})_{sl} = \sum_{i=1}^n f_i^{(s)} H_{il} = \sum_{i \in V_s} H_{il} - \frac{|V_s|}{n} \sum_{i=1}^n H_{il}$$
$$= \sum_{i \in V_s} H_{il} - \frac{|V_s|}{n}.$$

Thus, $(\boldsymbol{F}^\top \boldsymbol{H})_{sl} = 0$ if and only if $\sum_{i \in V_s} H_{il} = \frac{|V_s|}{n}$, which is exactly equal to the fairness condition in Eq. (1). $\square$

B. Shallow Fair NMF Model

We begin with a shallow Fair NMF (FNMF) that serves as the foundation for our deep hierarchical extension. To encode fairness, we augment the standard tri-factorization with a penalty derived from Lemma 1, encouraging proportional representation during optimization:

$$\min_{\boldsymbol{H}, \boldsymbol{W} \geq 0} \|\boldsymbol{A} - \boldsymbol{H}\boldsymbol{W}\boldsymbol{H}^\top\|_F^2 + \lambda \, \mathcal{R}(\boldsymbol{H}), \qquad (7)$$

where $\boldsymbol{H} \in \mathbb{R}_+^{n \times k}$ are soft cluster memberships and $\boldsymbol{W} \in \mathbb{R}_+^{k \times k}$ captures inter-cluster interactions.

a) Fairness penalty.: We enforce demographic balance via the smooth quadratic penalty (i.e. L_2 norm):

$$\mathcal{R}(\boldsymbol{H}) = \|\boldsymbol{F}^\top \boldsymbol{H}\|_F^2 = \sum_{s=1}^{m-1} \sum_{l=1}^k \left(\boldsymbol{f}^{(s)\top} \boldsymbol{h}_l\right)^2, \qquad (8)$$

where $\boldsymbol{f}^{(s)}$ is column s of $\boldsymbol{F}$ and $\boldsymbol{h}_l$ is column l of $\boldsymbol{H}$. By Lemma 1, $\mathcal{R}(\boldsymbol{H}) = 0$ iff the balance condition holds; decreasing $\mathcal{R}$ drives cluster compositions toward global group proportions.

Substituting (8) into (7) yields our shallow FNMF objective:

$$\min_{\boldsymbol{H}, \boldsymbol{W} \geq 0} \underbrace{\|\boldsymbol{A} - \boldsymbol{H}\boldsymbol{W}\boldsymbol{H}^\top\|_F^2}_{\text{utility term}} + \underbrace{\lambda \|\boldsymbol{F}^\top \boldsymbol{H}\|_F^2}_{\text{fairness term}}, \qquad (9)$$

with λ controlling the utility–fairness trade-off.

C. Deep Fair NMF (DFNMF) Model

Shallow tri-factorization is transparent and efficient, but can miss multi-level structure on large graphs. We therefore adopt a deep tri-factorization inspired by [29] in which successive nonnegative layers capture increasingly coarse communities. Let

$$\boldsymbol{\Psi} = \prod_{i=1}^p \boldsymbol{H}_i, \quad \boldsymbol{H}_i \in \mathbb{R}_+^{r_{i-1} \times r_i} \qquad (10)$$

where $r_0 = n \geq \cdots \geq r_p = k$. and $\boldsymbol{W}_p \in \mathbb{R}_+^{k \times k}$ encode final inter-/intra-cluster interactions. The graph is reconstructed as

$$\boldsymbol{A} \approx \boldsymbol{\Psi}\boldsymbol{W}_p\boldsymbol{\Psi}^\top \qquad (11)$$

yielding direct soft memberships (columns of $\boldsymbol{\Psi}$) without any post-hoc clustering.

a) Unified objective.: We integrate the balance penalty from Eq. (8) at the final layer:

$$\min_{\{\boldsymbol{H}_i\}_{i=1}^p \geq 0, \, \boldsymbol{W}_p \geq 0} \mathcal{L} = \underbrace{\|\boldsymbol{A} - \boldsymbol{\Psi}\boldsymbol{W}_p\boldsymbol{\Psi}^\top\|_F^2}_{\text{utility}} + \underbrace{\lambda \|\boldsymbol{F}^\top \boldsymbol{\Psi}\|_F^2}_{\text{fairness}},$$
$$(12)$$

The regularizer acts on $\boldsymbol{\Psi}$, i.e., the final memberships, and λ tunes the utility–fairness trade-off.

b) Pipeline illustration.: Figure 2 depicts the hierarchy $(\boldsymbol{H}_1, \boldsymbol{H}_2, \boldsymbol{H}_3)$ and a 45-node example with two sensitive groups (triangles/circles). Small λ preserves intrinsic structure but yields imbalanced clusters (e.g., 5:9, 5:11, 8:7); large λ adjusts memberships toward parity (e.g., 7:11, 5:7, 6:9). This demonstrates controllable movement along the utility–fairness spectrum by tuning λ.

D. Optimization

We solve the DFNMF objective in Eq. (12) with alternating minimization [30], updating each factor while holding the others fixed. As the problem is non-convex, this procedure converges to a local optimum but does not guarantee global optimality.

a) Pretraining (Algorithm 1, lines 2–6).: To accelerate convergence, we initialize each layer via sequential NMTF: first factorize $\boldsymbol{A} \approx \boldsymbol{H}_1\boldsymbol{W}_1\boldsymbol{H}_1^\top$, then recursively factorize $\boldsymbol{W}_{i-1} \approx \boldsymbol{H}_i\boldsymbol{W}_i\boldsymbol{H}_i^\top$ for $i = 2, \ldots, p$. This provides warm starts for $\{\boldsymbol{H}_i\}$ and $\boldsymbol{W}_p$, reducing wall time in practice [31].

b) Fine-tuning (Algorithm 1, lines 7–16).: With $\boldsymbol{\Psi}$ defined as in Eq. (10), we alternate updates for the membership blocks $\{\boldsymbol{H}_i\}$ and the interaction matrix $\boldsymbol{W}_p$.

1) Update rule for membership blocks $\boldsymbol{H}_i$: Fix all variables except $\boldsymbol{H}_i$. Using the block products

$$\boldsymbol{\Psi}_i = \boldsymbol{H}_1 \cdots \boldsymbol{H}_{i-1}, \qquad \boldsymbol{\Phi}_i = \boldsymbol{H}_{i+1} \cdots \boldsymbol{H}_p,$$

(with $\boldsymbol{\Psi}_1 = \boldsymbol{I}$ and $\boldsymbol{\Phi}_p = \boldsymbol{I}$), the subproblem is

$$\min_{\boldsymbol{H}_i \geq 0} \mathcal{L}(\boldsymbol{H}_i) = \|\boldsymbol{A} - \boldsymbol{\Psi}_i\boldsymbol{H}_i\boldsymbol{\Phi}_i\boldsymbol{W}_p\boldsymbol{\Phi}_i^\top\boldsymbol{H}_i^\top\boldsymbol{\Psi}_i^\top\|_F^2 \qquad (13)$$
$$+ \lambda \|\boldsymbol{F}^\top\boldsymbol{\Psi}_i\boldsymbol{H}_i\boldsymbol{\Phi}_i\|_F^2.$$

Following a standard Lee–Seung style multiplicative update derived [24] via KKT conditions (derivation omitted for brevity), we obtain

$$\boldsymbol{H}_i \leftarrow \boldsymbol{H}_i \odot \left(\frac{\mathcal{N}_i}{\mathcal{D}_i}\right)^{1/4}, \qquad (14)$$

where we introduce compact shorthands

$$\mathcal{N}_i = \boldsymbol{\Psi}_i^\top \left(\boldsymbol{A}^\top\boldsymbol{\Psi}\boldsymbol{W}_p + \boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{W}_p^\top + \lambda \left[\boldsymbol{F}\boldsymbol{F}^\top\boldsymbol{\Psi}\right]^-\right)\boldsymbol{\Phi}_i^\top,$$
$$\mathcal{D}_i = \boldsymbol{\Psi}_i^\top \left(\boldsymbol{\Psi}\boldsymbol{W}_p^\top\boldsymbol{\Psi}^\top\boldsymbol{\Psi}\boldsymbol{W}_p + \boldsymbol{\Psi}\boldsymbol{W}_p\boldsymbol{\Psi}^\top\boldsymbol{\Psi}\boldsymbol{W}_p^\top \right.$$
$$\left. + \lambda \left[\boldsymbol{F}\boldsymbol{F}^\top\boldsymbol{\Psi}\right]^+\right)\boldsymbol{\Phi}_i^\top,$$

and $\boldsymbol{B}^+ = \max(\boldsymbol{B}, 0)$, $\boldsymbol{B}^- = -\min(\boldsymbol{B}, 0)$ so that $\boldsymbol{B} = \boldsymbol{B}^+ - \boldsymbol{B}^-$. Update (14) is used in Algorithm 1, line 13. The detailed derivation is discussed in the Appendix VIII-B.
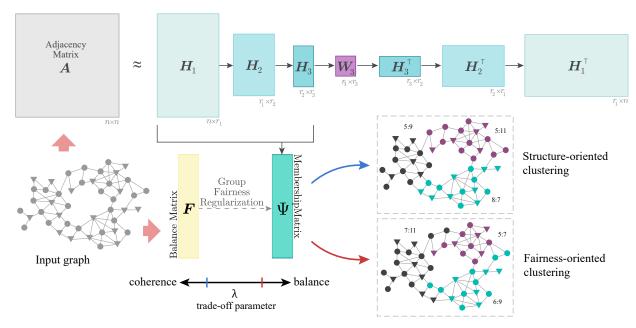
Fig. 2: **DFNMF schematic and example.** A 45-node graph with imbalanced gender distribution of 40%/60%(27 ○, 18 ▽) is factorized through $\boldsymbol{H}_1, \boldsymbol{H}_2, \boldsymbol{H}_3$. Two solutions illustrate the effect of λ: small λ preserves structure but yields imbalance (5:9, 5:11, 8:7); large λ improves parity (7:11, 5:7, 6:9), highlighting the utility–fairness trade-off.

Algorithm 1 [Balanced] Deep Fair NMF (DFNMF)

Input: The adjacency matrix of graph $\mathcal{G}$, $\boldsymbol{A}$; layer size of each layer, r_i; fairness regularization parameter λ;
Output: $\boldsymbol{W}_i$ $(1 \le i < p)$, $\boldsymbol{H}_i$ $(1 \le i < p)$, and the membership matrix $\boldsymbol{\Psi}$;

1: Constructing the balance fairness matrix $\boldsymbol{F}$ according to Definition 2;
2: ▷ **Pre-training process:**
3: $\boldsymbol{W}_1, \boldsymbol{H}_1 \leftarrow$ NMTF$(\boldsymbol{A}, r_1)$;
4: **for** $i = 2$ **to** p **do**
5: $\quad \boldsymbol{W}_i, \boldsymbol{H}_i \leftarrow$ NMTF$(\boldsymbol{W}_{i-1}, r_i)$;
6: **end for**
7: ▷ **Fine-tuning process:** following Section IV-C
8: $\boldsymbol{W}_i, \boldsymbol{H}_i \leftarrow$ DFNMF$(\boldsymbol{\Psi}_i, p)$; deep hierarchical learning
9: **while** convergence not reached **do**
10: $\quad$ **for** $i = 1$ **to** p **do**
11: $\quad\quad \boldsymbol{\Psi}_{i-1} \leftarrow \prod_{\tau=1}^{i-1} \boldsymbol{H}_\tau (\boldsymbol{\Psi}_0 \leftarrow \boldsymbol{I})$;
12: $\quad\quad \boldsymbol{\Phi}_{i+1} \leftarrow \prod_{\tau=i+1}^{p} \boldsymbol{H}_\tau (\boldsymbol{\Phi}_{p+1} \leftarrow \boldsymbol{I})$;
13: $\quad\quad$ Update clustering matrix $\boldsymbol{H}_i$ using (14);
14: $\quad\quad \boldsymbol{\Psi}_i \leftarrow \boldsymbol{\Psi}_{i-1} \boldsymbol{H}_i$;
15: $\quad\quad$ Update interaction matrix $\boldsymbol{W}_p$ using (16);
16: $\quad$ **end for**
17: **end while**

---

2) Update rule for interaction matrix $\boldsymbol{W}_p$: By fixing $\{\boldsymbol{H}_i\}_{i=1}^{p}$, we can update $\boldsymbol{W}_p$ by solving

$$\min_{\boldsymbol{W}_p} \mathcal{L}(\boldsymbol{W}_p) = \|\boldsymbol{A} - \boldsymbol{\Psi} \boldsymbol{W}_p \boldsymbol{\Psi}^\top\|_F^2, \quad \text{s.t.} \quad \boldsymbol{W}_p \ge 0, \quad (15)$$

Let $\bar{\boldsymbol{A}} = \boldsymbol{\Psi}^\top \boldsymbol{A} \boldsymbol{\Psi}$ and $\boldsymbol{S} = \boldsymbol{\Psi}^\top \boldsymbol{\Psi}$. The multiplicative update

becomes

$$\boldsymbol{W}_p \leftarrow \boldsymbol{W}_p \odot \frac{\bar{\boldsymbol{A}}}{\boldsymbol{S} \boldsymbol{W}_p \boldsymbol{S}}, \quad (16)$$

which corresponds to Algorithm 1, line 15.

a) Remarks.: (i) Eqs. (14)–(16) preserve nonnegativity and monotonically decrease the objective under the usual assumptions for multiplicative updates. (ii) The shorthands $\mathcal{N}_i, \mathcal{D}_i$ compactly collect terms as auxiliary variables improving readability.

E. Computational Complexity

Our method involves iterative updates of the factor matrices $\boldsymbol{W}_p$ (16) and $\boldsymbol{H}_i$ (14) until convergence. The update of $\boldsymbol{W}_p$ has a complexity of $\mathcal{O}(n^2 k + nk^2 + k^3)$, while the update of each $\boldsymbol{H}_i$, which incorporates the fairness regularization, has a complexity of $\mathcal{O}(n^2 k + nk^2 + n(m-1)k + nr_i k + r_i(m-1)k)$. In practice, k, m, and r_i are typically small, making the overall cost dominated by operations involving n. To further improve efficiency, our implementation uses sparse CSR representations of $\boldsymbol{A}$, reducing the dominant matrix operations to $\mathcal{O}(|E|k)$, where $|E|$ is the number of non-zero entries in $\boldsymbol{A}$ (i.e. number of edges). This yields an overall per-iteration complexity of $\mathcal{O}(p|E|k)$. Optional block coordinate descent (not used in our experiments) can further reduce memory from $\mathcal{O}(n^2)$ to $\mathcal{O}(bk)$.

V. EXPERIMENTS

This section presents experimental evaluations of our proposed model compared to baselines on both real-world and synthetic datasets. Results are assessed based on the utility and fairness of partitioning under various conditions.

A. Experimental Setup

All experiments were conducted on a NVIDIA A3090 GPU with 24 GB of memory. Sparse graph operations were implemented using the `scipy.sparse` classes of `csr_matrix` and `coo_matrix`, and their counterparts in `torch.sparse`. Optimized sparse implementations are detailed in code repository. We use a 500 epochs max, and $p = 4$ layers with 64 components (unless otherwise stated). The final layer is always projected to k, the number of communities (i.e., clusters) for DFNMF. The depth ($p = 4$) offers sufficient expressiveness while avoiding over-smoothing in deeper architectures [31]; deeper variants showed no empirical improvement. All results and ablations (Appendix 2 IX) average over 10 seeds; the $\lambda^\star$ selection rule was cross-checked via a linear scalarization to confirm robustness (see Section V-E).

1) Unsupervised Setting: As we address the unsupervised task of clustering, there is no train-validation-test split. Instead, the entire dataset is used as input during each model run. Importantly, no ground-truth labels are accessed during training or inference. Label-based metrics (e.g., ARI or accuracy) are used only *post hoc* for evaluation assessment.

2) Parameter Selection: We sweep λ on a logarithmic grid $\{10^{-3}, \ldots, 10^3\}$ and form the utility–fairness curve $(\tilde{Q}(\lambda), \tilde{B}(\lambda))$ after min–max scaling of Q and $\bar{B}$ to $[0, 1]$ (per dataset and k). We first retain the *Pareto front* Λ_P (undominated points), then select

$$\lambda^\star = \arg\min_{\lambda \in \Lambda_\mathrm{P}} \left\| (\tilde{Q}(\lambda), \tilde{B}(\lambda)) - (1, 1) \right\|_2,$$

with a tie–breaker favoring smaller $|\tilde{Q}(\lambda) - \tilde{B}(\lambda)|$ (closer to the identity guide as shown in Section V-E4). This setting obtains a robust and reproducible $\lambda^\star$ scheme, which are reported together with their bracket $\{\lambda^\star/10, 10\lambda^\star\}$ in the appendix B. All the experimental results throughout the paper are reported according to our optimal $\lambda^\star$ for each dataset.

3) Baselines: We compare DFNMF with seven baselines introduced previously in Section II. They comprise vanilla models of spectral clustering (SC) [32], tri-factor NMF (NMTF) [27], GNN-based model (DMoN) [23], and fairness-aware models: fair, scalable, individual spectral models (FairSC, sFSC, iFSC) [14]–[16], and individually fair NMTF (iFNMTF) [17]. All hyperparameters are tuned according to the original papers.

4) Evaluation Measures: We assess clustering *utility* on labeled datasets using **accuracy (ACC)** and **adjusted Rand index (ARI)**, and on unlabeled datasets using Newman's **modularity (Q)** [33]. Accuracy calculates the proportion of nodes correctly assigned to their respective clusters. ARI is a pairwise agreement measure between predicted and ground-truth clusterings, adjusted for chance, and ranges from -1 (random clustering), through 0 (chance-level agreement), to 1 (perfect clustering). Modularity quantifies the strength of division of a graph into clusters by comparing the observed intra-cluster connectivity to that expected under a random null model, also ranging from -1 to 1. For *fairness*, we report **average balance** ($\bar{B}$) [14], [16] and **statistical parity deviation** (Δ_{SP}) [34]. $\bar{B}$ computes the mean minimum group proportion across clusters, where higher values in $[0, 1]$ indicate fairer clusters. Δ_{SP} measures deviation from global group proportions within each cluster:

$$\Delta_{SP} = \frac{1}{k} \sum_{l=1}^{k} \sum_{s=1}^{m} \left| \frac{|V_s \cap C_l|}{|C_l|} - \frac{|V_s|}{|V|} \right|, \qquad (17)$$

where lower values (approaching 0) imply better demographic parity. All metrics are averaged over 10 runs. ACC and $\bar{B}$ lie in $[0, 1]$, while ARI, Q, and Δ_{SP} range in $[-1, 1]$.

B. Datasets

We evaluate DFNMF on 11 networks (8 real-world, 3 synthetic) with diverse structural setups, group imbalances, and homophily levels. Dataset characteristics, including number of nodes ($|V|$), edges ($|E|$), sensitive groups, number of clusters (k), edge density, and homophily, are summarized in Table II.

Homophily quantifies the tendency of nodes to connect within the same sensitive group, indicating inherent network bias. Thus, datasets with higher homophily tend to exacerbate the impact of demographic imbalances on fairness.

TABLE II: Summary statistics for datasets used in experiments.

Network	$|V|$	$|E|$	Attribute (# groups)	Edge Density	Homophily	k
SBM	2,000	267,430	attr (2)	0.133	0.82	5
	5,000	978,959	attr (2)	0.078	0.82	5
	10,000	2,603,190	attr (2)	0.052	0.82	5
Pokec-n	67,797	882,765	age (4)	0.0384	0.399	2
Pokec-z	66,570	729,129	age (4)	0.0329	0.360	2
NBA	403	8,285	ethnicity (2)	0.102	0.72	2
Diaries	120	348	gender (2)	0.048	0.61	–
Friendship	134	406	gender (2)	0.049	0.60	–
Facebook	156	1,437	gender (2)	0.120	0.57	–
DrugNet	293	284	ethnicity (3)	0.014	0.88	–
LastFM	7,624	27,806	country (6)	0.001	0.92	–

1) Synthetic Networks: are generated using an extended Stochastic Block Model (**SBM**) following [14]–[16], with explicitly controlled group memberships and cluster assignments. Each node set $V=[n]$ is partitioned into m groups $V = V_1 \dot{\cup} \ldots \dot{\cup} V_m$ and k clusters $V = C_1 \dot{\cup} \ldots \dot{\cup} C_k$, ensuring proportional representation of each sensitive group within clusters (fair clustering by construction).

2) Real-World Networks: Our collection of real datasets represents diverse social and interaction graphs with varying demographics and biases [35]–[38]: small-scale datasets (*Facebook*, *Friendship*, *Diaries*) exhibit moderate gender imbalance (around 60% majority groups), and medium-sized datasets like *DrugNet* and *NBA* show stronger imbalances (majority ethnic groups above 70%). Larger datasets such as *Pokec-n* and *Pokec-z* present considerable imbalance (majority age group over 70%) with moderate homophily (0.36–0.40), while highly sparse and homophilous graphs like *LastFM* (country, homophily = 0.92) and *DrugNet* (ethnicity, homophily = 0.88) pose greater fairness challenges. These variations allow comprehensive exploration of fairness-utility trade-offs.
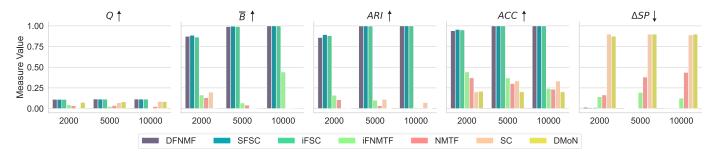
Fig. 3: SBM networks with varying node sizes: comparison of clustering and fairness metrics. Arrows (↑/↓) indicate whether higher/lower is better.

C. Results on Synthetic Datasets

To benchmark the baselines under controlled conditions, we conducted experiments on three variants of the SBM networks with 2,000, 5,000, and 10,000 nodes. Results are depicted in Figure 3. SFSC and FSC baselines report identical results; thus, only SFSC is visualized to conserve space. The results indicate that DFNMF, SFSC, and iFSC consistently outperform other baselines, including vanilla models (SC, NMTF, DMoN) and the iFNMTF by reporting the highest Q, $\overline{B}$, ARI, ACC, and lowest ΔSP. These findings strongly support that the three models have an advantage in trading-off clustering utility and fairness under varying degrees of structural complexity and group imbalance on SBM networks.

D. Results on Real-world Datasets

We further evaluate DFNMF and baseline models on a variety of real-world networks, with diverse structural and fairness challenges. These include datasets with strong group imbalance (e.g., Pokec, NBA), as well as sparse, highly homophilous graphs such as *LastFM* and *DrugNet*, as characterized in Section V-B and Table II. Given the comparable performance of spectral methods on synthetic data, we focus here on their limitations in more complex, real-world scenarios.

1) UnLabeled Datasets: Table III presents results on datasets without ground-truth labels, evaluated using average balance ($\overline{B}$) and modularity (Q) under two cluster settings (C=5 and C=10). On *DrugNet* and *LastFM*, DFNMF shows a clear advantage, achieving both high modularity and fairness. In contrast, spectral baselines (SFSC, FSC, iFSC) perform poorly due to their rigid fairness constraints, which often fail to adapt to highly sparse and biased network structures. On smaller networks such as *Facebook*, *Friendship*, and *Diaries*, DFNMF maintains top fairness performance and ranks second in modularity—slightly behind SFSC or iFSC—demonstrating strong overall adaptability even in simpler settings.

2) Labeled Datasets: We evaluate DFNMF against SFSC—a fair spectral clustering baseline—and DMoN, a state-of-the-art GNN-based model. This comparison is designed to assess DFNMF's scalability and fairness performance across varying data complexities. Specifically, we test whether spectral methods like SFSC degrade under high-dimensional settings, and whether DMoN's prior limitations are due to dataset scale or inherent model bias. Table IV reports results on three Labeled

TABLE III: Performance on unLabeled datasets for k=5 and k=10 number of clusters.

Model		DrugNet		Diaries		Facebook		Friendship		LastFM	
		k=5	k=10	k=5	k=10	k=5	k=10	k=5	k=10	k=5	k=10
DMoN	$\overline{B}$	0.00	0.00	0.263	0.034	0.267	0.194	0.182	0.040	0.00	0.00
	Q	0.326	0.324	0.145	0.165	0.047	0.034	0.129	0.121	0.526	0.360
SC	$\overline{B}$	0.030	0.020	0.681	0.467	0.295	0.181	0.362	0.282	0.005	0.002
	Q	**0.607**	**0.633**	0.631	0.669	0.454	0.450	0.575	0.602	0.418	0.419
SFSC	$\overline{B}$	0.052	0.026	0.684	0.494	0.601	0.436	0.485	0.399	0.067	0.033
	Q	27.00	52.56	**0.808**	**0.698**	0.500	0.512	0.626	0.665	0.034	0.040
FSC	$\overline{B}$	0.052	0.026	0.684	0.494	0.601	0.436	0.485	0.399	0.067	0.033
	Q	0.270	0.525	0.808	0.698	0.500	**0.512**	0.626	0.665	0.034	0.040
iFSC	$\overline{B}$	0.055	0.037	**0.800**	0.495	0.564	0.452	0.581	0.520	0.065	0.032
	Q	0.279	0.457	0.657	0.697	**0.515**	0.496	0.624	**0.683**	0.007	0.013
iFNMF	$\overline{B}$	0.098	0.114	0.706	0.578	0.527	0.001	0.613	0.551	0.071	0.029
	Q	0.116	0.058	0.238	0.129	0.239	0.004	0.216	0.621	0.254	0.225
DFNMF	$\overline{B}$	**0.162**	**0.141**	0.787	**0.707**	**0.767**	**0.614**	**0.614**	**0.623**	**0.091**	**0.084**
	Q	0.591	0.524	0.716	0.594	0.503	0.432	**0.666**	0.604	**0.420**	**0.455**

TABLE IV: Performance (**mean** ± **std**) on Labeled datasets. Arrows (↑/↓) show if higher/lower is better.

Metric	Model	Pokec-n	Pokec-z	NBA
ARI ↑	DFNMF	**0.0051** ± **0.001**	**0.0158** ± **0.002**	**0.1203** ± **0.012**
	SFSC	0.0009 ± 0.000	0.0009 ± 0.000	0.0825 ± 0.007
	DMoN	0.0024 ± 0.000	0.0022 ± 0.000	0.0741 ± 0.006
Q ↑	DFNMF	**0.2067** ± **0.005**	**0.1910** ± **0.026**	**0.1344** ± **0.012**
	SFSC	0.0001 ± 0.000	0.0001 ± 0.000	0.1089 ± 0.013
	DMoN	0.1801 ± 0.003	0.1625 ± 0.004	0.1162 ± 0.008
ACC ↑	DFNMF	0.6879 ± 0.052	0.7767 ± 0.096	**0.6765** ± **0.010**
	SFSC	**0.8476** ± **0.012**	**0.8165** ± **0.042**	0.6500 ± 0.004
	DMoN	0.5252 ± 0.035	0.5062 ± 0.039	0.5232 ± 0.026
$\overline{B}$ ↑	DFNMF	**0.1844** ± **0.017**	**0.2905** ± **0.025**	**0.4373** ± **0.028**
	SFSC	0.0974 ± 0.010	0.2249 ± 0.021	0.3590 ± 0.052
	DMoN	0.1135 ± 0.002	0.0589 ± 0.004	0.0012 ± 0.040
Δ_{SP} ↓	DFNMF	**0.0194** ± **0.002**	**0.0189** ± **0.000**	**0.0018** ± **0.000**
	SFSC	0.5664 ± 0.042	0.0573 ± 0.016	0.0030 ± 0.000
	DMoN	0.3856 ± 0.020	0.4143 ± 0.053	0.0087 ± 0.001

datasets—*Pokec-n*, *Pokec-z*, and *NBA*—using both utility (ACC, ARI, Q) and fairness metrics ($\overline{B}$, Δ_{SP}). DFNMF consistently achieves the best overall performance, offering a strong balance

between clustering quality and fairness. On Pokec datasets, it improves Q and ARI while also ensuring $\overline{B}$ and lower Δ_{SP}. SFSC, while competitive in raw accuracy, fails on fairness and modularity due to its rigid constraints that tend to favor trial solutions with dominant groups. DMoN shows moderate utility but poor fairness, likely due to bias propagation through GNN embeddings and lack of clear debiasing policies. On the NBA dataset, which presents moderate homophily and high group imbalance, DFNMF again outperforms both baselines across all metrics, demonstrating its adaptability and robustness in handling complex group structures while maintaining fairness.

E. Convergence, Interpretability, scalability, and Fairness-Utility Trade-offs analysis

1) Convergence Analysis.: DFNMF employs a two-phase optimization strategy: a multi-layer pretraining phase with shallow NMTF (with known convergence guarantees [39]) and a deep fine-tuning phase minimizing the fairness-aware objective (12). Figure 4 shows convergence curves on synthetic graphs (5K and 10K nodes) for both the fairness-aware variant ($\lambda{=}10$) and a fairness-agnostic version ($\lambda{=}0$). Both converge within 150 iterations; the fairness-regularized version stabilizes at a slightly higher loss, reflecting the additive influence of fairness penalties.
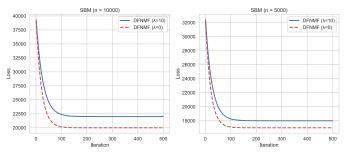

Fig. 4: Convergence curves on SBM graphs (5K and 10K nodes), comparing DFNMF with/without fairness regularization ($\lambda = 10$ vs. $\lambda = 0$). Both variants converge rapidly; the fair version yields a higher loss due to the added fairness penalty.

2) Interpretability Analysis: DFNMF is interpretable by construction: its nonnegative factors encode parts-based *soft memberships* across the hierarchy of nonnegative affinities. On the 60-node graph in Fig. 5, the first layer assigns nodes to 12 micro-clusters (H_1), the second maps micro-clusters to three communities (H_2), and $\Psi = H_1 H_2$ yields final node–community affinities.

A compact slice of H_1 (nodes 1–5 and 56–60) is shown in Table V; the full H_1 appears in Appendix IX-C. Figure 6 illustrates the relationship of micro-cluster→community mapping H_2 with the corresponding (trimmed) node→community matrix Ψ; complete matrices are deferred to Appendix IX-C.

Two patterns underpin interpretability. *(i) Sparse micro-memberships:* rows of H_1 are typically few-peaked, indicating that most nodes participate in a small number of micro-clusters; rows with multiple peaks (e.g., node 5 in Table V) flag potential

TABLE V: Node → micro-cluster memberships (H_1). We show nodes 1–5 and 56–60; full table in Appendix IX-C.

Node	Micro-clusters											
	A	B	C	D	E	F	G	H	I	J	K	L
1	0.26	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.01	0.00	0.00	0.00
2	0.25	0.00	0.01	0.00	0.01	0.09	0.02	0.00	0.01	0.00	0.04	0.00
3	0.21	0.00	0.01	0.00	0.09	0.03	0.00	0.00	0.00	0.00	0.11	0.00
4	0.23	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00
5	0.31	0.00	0.02	0.00	0.01	0.15	0.05	0.10	0.05	0.02	0.09	0.00
56	0.01	0.00	0.24	0.00	0.00	0.04	0.13	0.00	0.01	0.00	0.25	0.00
57	0.06	0.00	0.26	0.00	0.00	0.05	0.01	0.09	0.06	0.01	0.26	0.00
58	0.00	0.00	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.24	0.00
59	0.00	0.00	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00
60	0.00	0.00	0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00

bridge nodes. *(ii) Structured aggregation:* H_2 is nearly one-hot for core micro-clusters (e.g., A→I, B→II, C→III), while boundary micro-clusters (e.g., D/E/H) spread mass across communities—capturing inter-community conduits. Multiplying H_1 by H_2 consolidates these signals: nodes with concentrated micro-membership become near one-hot in Ψ, whereas mixed rows remain soft. This traceability—from node→micro (H_1) to micro→community (H_2) to node→community (Ψ)—supports transparent auditing of how local structure (under fairness regularization) aggregates into global communities.

3) Scalability Analysis: Table VI reports wall-clock runtime (seconds) and memory footprint of DFNMF against baselines on synthetic Erdős-Rényi graphs with varying sizes, averaged over 10 runs on a single NVIDIA A3090 GPU.

TABLE VI: Scalability on Erdős-Rényi Graphs ($p = 2$, $k = 128$)

Scale		Methods				
Nodes	Edges	sFSC	iFSC	NMTF	DMoN	DFNMF
		Runtime (seconds)				
10^4	$\simeq 10^5$	42.3	48.7	2.8	8.3	**1.5**
10^5	$\simeq 10^6$	1842	2103	31.6	78.4	**9.8**
10^6	$\simeq 10^7$	OOM	OOM	413	8932	**92.4**
		Memory (GB)				
10^4	$\simeq 10^5$	2.1	2.4	0.8	1.2	**0.4**
10^5	$\simeq 10^6$	45.2	52.3	5.4	8.1	**2.1**
10^6	$\simeq 10^7$	OOM	OOM	21.3	OOM	**12.1**

DFNMF demonstrates higher scalability with near-linear runtime growth $O(|E|kp)$ versus spectral methods' $O(n^3)$. Beyond 100K nodes, spectral approaches exhaust memory while DFNMF processes million-node graphs efficiently. The advantage over NMTF is due to CSR-based sparse operations and pre-training process that remarkably reduces iterations.

4) Fairness–Utility Trade-offs: We assess trade-offs similar to [40], [41] at $k{=}5$ on *DrugNet* and *LastFM* (Fig. 7) by sweeping $\lambda \in [10^{-3}, 10^3]$ and plotting $(Q, \overline{B})$. Blue points are DFNMF solutions; other markers are baselines. We first retain the *Pareto front* (undominated DFNMF points) and then select the *ideal-point* configuration (closest to $(1, 1)$ after min–max
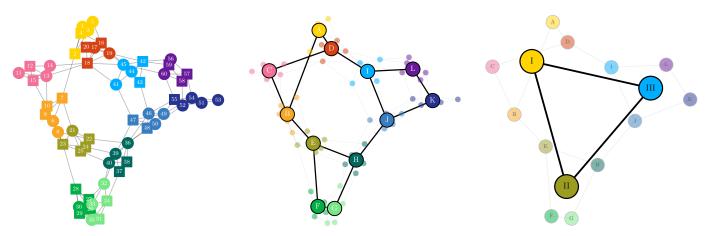
Fig. 5: **DFNMF hierarchy on a 60-node graph.** (a) Input graph; node shapes denote sensitive groups. (b) Micro-clusters (A–L) discovered by the first layer (H_1). (c) Three coarse communities obtained by aggregating micro-clusters via H_1; final node–community affinities are $\Psi = H_1 H_2$.

H_2 (A–L → I–III)				Ψ (nodes $\to$ $clusters$)			
Micro cluster	Clusters			Node	Clusters		
	I	II	III		I	II	III
A	1.04	0.00	0.00	1	0.27	0.00	0.00
B	0.00	1.00	0.00	2	0.26	0.00	0.01
C	0.00	0.00	1.00	3	0.22	0.00	0.03
D	0.00	0.05	0.01	4	0.24	0.00	0.00
E	0.02	0.00	0.19	5	0.33	0.01	0.02
F	0.00	0.00	0.00			$\vdots$	
G	0.08	0.00	0.00				
H	0.06	0.08	0.00	56	0.02	0.00	0.24
I	0.00	0.00	0.00	57	0.07	0.01	0.26
J	0.00	0.02	0.06	58	0.00	0.00	0.22
K	0.00	0.00	0.01	59	0.00	0.00	0.31
L	0.02	0.00	0.01	60	0.00	0.00	0.19

Fig. 6: Micro→community (H_2) and node→community ($\Psi = H_1 H_2$) soft memberships.

scaling), shown as the green star at $\lambda^\star = 100$—the same setting reported in Table III. Dashed identity lines provide a balanced trade-off guide (top-right is best), and shaded curvature indicates empirical fronts.

Consistent patterns emerge: DMoN and SC attain higher Q but poorer balance; fairness-oriented spectral baselines (FSC/SFSC/iFSC) improve $\bar{B}$ at the expense of Q. In contrast, the DFNMF sweep *spans the spectrum*: increasing λ moves solutions rightward (higher $\bar{B}$) with some loss in Q, while decreasing λ moves them upward (higher Q) with lower $\bar{B}$. At the extremes, DFNMF reaches fronts that surpass baselines on either objective individually, while $\lambda^\star$ lies near the identity guide, offering a balanced, high-quality operating point.

VI. CONCLUSION AND OUTLOOK

We introduce DFNMF, an end-to-end deep matrix factorization framework that directly integrates fairness constraints into graph clustering. Unlike existing approaches that rely on rigid
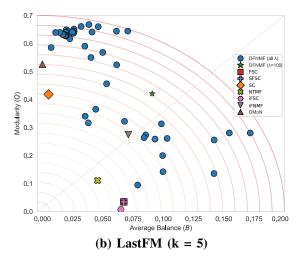

(a) **DrugNet (k = 5)**

(b) **LastFM (k = 5)**

Fig. 7: Pareto plots for $k=5$. Blue: DFNMF across $\lambda \in [10^{-3}, 10^3]$; green star: $\lambda^\star = 100$ selected on the Pareto front via the ideal-point rule. Dashed identity lines mark balanced trade-offs; shaded curves indicate empirical fronts.

constraints or multi-stage pipelines, DFNMF enables flexible fairness-utility trade-offs through a single parameter while maintaining computational efficiency with near-linear scaling in graph edges. The nonnegative factorization provides inherent interpretability through parts-based decomposition, addressing a key limitation of spectral methods.

Our theoretical analysis establishes formal connections between matrix-based regularization and demographic balance constraints, providing principled foundations for the approach. Comprehensive experiments across synthetic and real networks demonstrate that DFNMF consistently achieves superior fairness-utility trade-offs, often dominating state-of-the-art methods on the Pareto front while maintaining scalability to large graphs through sparse matrix operations.

Several promising directions emerge from this work. *Individual fairness* integration would ensure equitable treatment of similar nodes beyond group-level parity. *Capacity constraints* could maintain balanced partition sizes while preserving structural coherence. While our linear formulation offers interpretability, *neural extensions* could capture more complex combinatorial patterns—particularly through fair neural matrix factorization variants that maintain the end-to-end optimization benefits.

Broadly, it demonstrates the potential for principled fairness integration in graph learning tasks, opening avenues for fair community detection in sensitive domains such as healthcare networks, financial systems, and social platforms where demographic balance is algorithmically and socially critical.

VII. ACKNOWLEDGEMENTS

REFERENCES

[1] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis *et al.*, "Bias in data-driven artificial intelligence systems: An introductory survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1356, 2020.

[2] T. L. Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi, "A survey on datasets for fairness-aware machine learning," *WIREs Data Mining Knowl. Discov.*, vol. 12, no. 3, 2022.

[3] X. Zhao, S. Fabbrizzi, P. R. Lobo, S. Ghodsi, K. Broelemann, S. Staab, and G. Kasneci, "Adversarial reweighting guided by wasserstein distance to achieve demographic parity," in *IEEE Big Data*. IEEE, 2024, pp. 1605–1614.

[4] S. Tahmasebi, E. Müller-Budack, and R. Ewerth, "Verifying cross-modal entity consistency in news using vision-language models," in *ECIR (4)*, ser. Lecture Notes in Computer Science, vol. 15575. Springer, 2025, pp. 339–354.

[5] Y. Dong, J. Ma, S. Wang, C. Chen, and J. Li, "Fairness in graph mining: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 10, pp. 10 583–10 602, 2023.

[6] J. M. Álvarez, A. B. Colmenarejo, A. Elobaid, S. Fabbrizzi, M. Fahimi, A. Ferrara, S. Ghodsi, C. Mougan, I. Papageorgiou, P. R. Lobo, M. Russo, K. M. Scott, L. State, X. Zhao, and S. Ruggieri, "Policy advice and best practices on bias and fairness in AI," *Ethics Inf. Technol.*, vol. 26, no. 2, p. 31, 2024.

[7] W. Ju, S. Yi, Y. Wang, Q. Long, J. Luo, Z. Xiao, and M. Zhang, "A survey of data-efficient graph learning," in *IJCAI*, 2024, pp. 8104–8113.

[8] T. L. Quy, G. Friege, and E. Ntoutsi, "Multi-fair capacitated students-topics grouping problem," in *PAKDD (1)*, ser. LNCS, vol. 13935. Springer, 2023, pp. 507–519.

[9] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii, "Fair clustering through fairlets," in *NeurIPS*, 2017, pp. 5029–5037.

[10] M. Bateni, V. Cohen-Addad, A. Epasto, and S. Lattanzi, "A scalable algorithm for individually fair k-means clustering," in *AISTATS*, vol. 238. PMLR, 2024, pp. 3151–3159.

[11] Y. Dong, J. Kang, H. Tong, and J. Li, "Individual fairness for graph neural networks: A ranking based approach," in *KDD*. ACM, 2021, pp. 300–310.

[12] A. Chen, R. A. Rossi, N. Park, P. Trivedi, Y. Wang, T. Yu, S. Kim, F. Dernoncourt, and N. K. Ahmed, "Fairness-aware graph neural networks: A survey," *ACM Trans. Knowl. Discov. Data*, vol. 18, no. 6, pp. 138:1–138:23, 2024.

[13] C. Yang, J. Liu, Y. Yan, and C. Shi, "Fairsin: Achieving fairness in graph neural networks through sensitive information neutralization," in *AAAI*. AAAI Press, 2024, pp. 9241–9249.

[14] M. Kleindessner, S. Samadi, P. Awasthi, and J. Morgenstern, "Guarantees for spectral clustering with fairness constraints," in *ICML*, vol. 97, 2019, pp. 3458–3467.

[15] S. Gupta and A. Dukkipati, "Consistency of constrained spectral clustering under graph induced fair planted partitions," in *NeurIPS*, 2022, pp. 13 527–13 540.

[16] J. Wang, D. Lu, I. Davidson, and Z. Bai, "Scalable spectral clustering with group fairness constraints," in *AISTATS*, 2023, pp. 6613–6629.

[17] S. Ghodsi, S. A. Seyedi, and E. Ntoutsi, "Towards cohesion-fairness harmony: Contrastive regularization in individual fair graph clustering," in *PAKDD (1)*, vol. 14645, 2024, pp. 284–296.

[18] S. Wang, L. Hu, Y. Wang, X. He, Q. Z. Sheng, M. A. Orgun, L. Cao, F. Ricci, and P. S. Yu, "Graph learning based recommender systems: A review," in *Proceedings of the 30th IJCAI*, 2021, pp. 4644–4652.

[19] S. Ghodsi and E. Ntoutsi, "Affinity clustering framework for data debiasing using pairwise distribution discrepancy," in *EWAF*, ser. CEUR Proceedings, vol. 3442, 2023.

[20] S. Ghodsi, H. Alani, and E. Ntoutsi, "Context matters for fairness - a case study on the effect of spatial distribution shifts," *CoRR*, vol. abs/2206.11436, 2022.

[21] J. Dickerson, S. A. Esmaeili, J. Morgenstern, and C. J. Zhang, "SoK: Fair Clustering: Critique, Caveats, and Future Directions," in *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE Computer Society, Apr. 2025, pp. 698–713.

[22] J. Li, Y. Wang, and A. Merchant, "Spectral normalized-cut graph partitioning with fairness constraints," in *ECAI*, ser. Frontiers in Artificial Intelligence and Applications, vol. 372. IOS Press, 2023, pp. 1389–1397.

[23] A. Tsitsulin, J. Palowitch, B. Perozzi, and E. Müller, "Graph clustering with graph neural networks," *J. Mach. Learn. Res.*, vol. 24, pp. 127:1–127:21, 2023.

[24] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[25] N. Gillis, *Nonnegative matrix factorization*. SIAM, 2020.

[26] S. A. Seyedi, S. S. Ghodsi, F. A. Tab, M. Jalili, and P. Moradi, "Self-paced multi-label learning with diversity," in *ACML*, ser. Proceedings of Machine Learning Research, vol. 101. PMLR, 2019, pp. 790–805.

[27] D. Kuang, H. Park, and C. H. Q. Ding, "Symmetric nonnegative matrix factorization for graph clustering," in *SDM*, 2012, pp. 106–117.

[28] Y. Pei, N. Chakraborty, and K. P. Sycara, "Nonnegative matrix tri-factorization with graph regularization for community detection in social networks," in *IJCAI*. AAAI Press, 2015, pp. 2083–2089.

[29] A. Hajiveiseh, S. A. Seyedi, and F. A. Tab, "Deep asymmetric nonnegative matrix factorization for graph clustering," *Pattern Recognit.*, vol. 148, p. 110179, 2024.

[30] G. Trigeorgis, K. Bousmalis, S. Zafeiriou, and B. W. Schuller, "A deep semi-nmf model for learning hidden representations," in *ICML*, vol. 32. JMLR.org, 2014, pp. 1692–1700.

[31] P. D. Handschutter, N. Gillis, and X. Siebert, "A survey on deep matrix factorizations," *Comput. Sci. Rev.*, vol. 42, p. 100423, 2021.

[32] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[33] T. Chakraborty, A. Dalmia, A. Mukherjee, and N. Ganguly, "Metrics for community analysis: A survey," *ACM Computing Surveys*, vol. 50, no. 4, pp. 1–37, 2017.

[34] S. Verma and J. Rubin, "Fairness definitions explained," in *FairWare@ICSE*. ACM, 2018, pp. 1–7.

[35] E. Dai and S. Wang, "Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information," in *WSDM*, 2021, pp. 680–688.

[36] B. Rozemberczki and R. Sarkar, "Characteristic functions on graphs: Birds of a feather, from statistical descriptors to parametric models," in *CIKM*, 2020, pp. 1325–1334.

[37] M. R. Weeks, S. Clair, S. P. Borgatti, K. Radda, and J. J. Schensul, "Social networks of drug users in high-risk sites: Finding the connections," *AIDS and Behavior*, vol. 6, pp. 193–206, 2002.

[38] R. Mastrandrea, J. Fournet, and A. Barrat, "Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys," *PloS one*, vol. 10, no. 9, p. e0136497, 2015.

[39] F. Wang, T. Li, X. Wang, S. Zhu, and C. H. Q. Ding, "Community discovery using nonnegative matrix factorization," *Data Min. Knowl. Discov.*, vol. 22, no. 3, pp. 493–521, 2011.

[40] S. Tahmasebi, P. Moradi, S. Ghodsi, and A. Abdollahpouri, "An ideal point based many-objective optimization for community detection of complex networks," *Inf. Sci.*, vol. 502, pp. 125–145, 2019.

[41] S. Ghodsi, S. Tahmasebi, M. Jalili, and P. Moradi, "Many-objective evolutionary optimization using density peaks scoring selection strategy," in *GECCO Companion*. ACM, 2024, pp. 331–334.

[42] C. Lin, "On the convergence of multiplicative update algorithms for nonnegative matrix factorization," *IEEE Trans. Neural Networks*, vol. 18, no. 6, pp. 1589–1596, 2007.

[43] D. Xu, C. Ruan, E. Körpeoglu, S. Kumar, and K. Achan, "Rethinking neural vs. matrix-factorization collaborative filtering: the theoretical perspectives," in *ICML*, vol. 139, 2021, pp. 11 514–11 524.

[44] S. Rendle, W. Krichene, L. Zhang, and J. R. Anderson, "Neural collaborative filtering vs. matrix factorization revisited," in *RecSys*. ACM, 2020, pp. 240–248.

[45] A. Roy, C. Koutlis, S. Papadopoulos, and E. Ntoutsi, "Fairbranch: Mitigating bias transfer in fair multi-task learning," in *IJCNN*. IEEE, 2024, pp. 1–8.

[46] A. Roy, J. Horstmann, and E. Ntoutsi, "Multi-dimensional discrimination in law and machine learning - A comparative overview," in *FAccT*. ACM, 2023, pp. 89–100.

[47] S. Swati, A. Roy, E. Panagiotou, and E. Ntoutsi, "Mmm-fair: An interactive toolkit for exploring and operationalizing multi-fairness trade-offs," *CoRR*, vol. abs/2509.08156, 2025.

A. Epistemic Comparison of NMF & DNN

In this section, we discuss theoretical and epistemic differences between NMF-based and deep neural network (DNN)-based clustering, focusing on comparative strengths, interpretability, and underlying philosophies.

NMF models naturally facilitate unsupervised and supervised scenarios, well-suited for multi-layer optimization. Their fundamental assumption—objects perceived as additive combinations of meaningful parts—aligns closely with human perception mechanisms [24]. Nonnegative constraints further ensure interpretability; negative contributions are often meaningless in real-world tasks like face, image, or gene data analysis. Such nonnegative decompositions typically yield localized, semantically interpretable features (e.g., facial parts). Additionally, the inherent sparseness of NMF enhances representation, clearly distinguishing these models from purely distributed approaches.

While shallow NMFs are equivalent to single-layer perceptrons, differences emerge in multilayer architectures: NMFs remain linear reconstruction models, optimized by specialized multiplicative update rules [42], unlike nonlinear DNNs optimized via gradient-based chain rules. This distinction is not about determining a universally best method. Indeed, many successful DNN architectures utilize dot-product similarity mechanisms, alike factorization methods. Recent works indicate that NMF-based factorization models often provide interpretability and computational advantages in embedding-based tasks such as collaborative filtering [43], [44]. Consequently, emerging research increasingly blends NMF and DNN or extends NMF to deep hierarchical models to leverage complementary strengths and optimize application-specific trade-offs.

B. Derivation of update formula

To calculate the gradient of the objective function in Eq. (13), we first need to express the function as a trace expression. Then, we can solve Eq. (13) by introducing a Lagrangian multiplier matrix $\boldsymbol{\Theta}_i$ to ensure the nonnegativity constraints on $\boldsymbol{H}_i$. This results in an equivalent objective function as follows:

$$
\min_{\boldsymbol{H}_i, \boldsymbol{\Theta}_i} \mathcal{L}(\boldsymbol{H}_i, \boldsymbol{\Theta}_i) = \mathrm{Tr}(-2\boldsymbol{A}^\top \boldsymbol{\Psi}_i \boldsymbol{H}_i \boldsymbol{\Phi}_i \boldsymbol{W}_p \boldsymbol{\Phi}_i^\top \boldsymbol{H}_i^\top \boldsymbol{\Psi}_i^\top
$$
$$
+ \boldsymbol{\Psi}_i \boldsymbol{H}_i \boldsymbol{\Phi}_i \boldsymbol{W}_p^\top \boldsymbol{\Phi}_i^\top \boldsymbol{H}_i^\top \boldsymbol{\Psi}_i^\top \boldsymbol{\Psi}_i \boldsymbol{H}_i \boldsymbol{\Phi}_i \boldsymbol{W}_p \boldsymbol{\Phi}_i^\top \boldsymbol{H}_i^\top \boldsymbol{\Psi}_i^\top)
$$
$$
+ \lambda \mathrm{Tr}(\boldsymbol{\Phi}_i^\top \boldsymbol{H}_i^\top \boldsymbol{\Psi}_i^\top \boldsymbol{F} \boldsymbol{F}^\top \boldsymbol{\Psi}_i \boldsymbol{H}_i \boldsymbol{\Phi}_i) - \mathrm{Tr}(\boldsymbol{\Theta}_i \boldsymbol{H}_i^\top). \quad (18)
$$

By setting the partial derivative of $\mathcal{L}(\boldsymbol{H}_i, \boldsymbol{\Theta}_i)$ with respect to $\boldsymbol{H}_i$ to $\boldsymbol{0}$, we have:

$$
\boldsymbol{\Theta}_i = -2\boldsymbol{\Psi}_i^\top \boldsymbol{A}^\top \boldsymbol{\Psi} \boldsymbol{W}_p \boldsymbol{\Phi}_i^\top - 2\boldsymbol{\Psi}_i^\top \boldsymbol{A} \boldsymbol{\Psi} \boldsymbol{W}_p^\top \boldsymbol{\Phi}_i^\top
$$
$$
+ 2\boldsymbol{\Psi}_i^\top \boldsymbol{\Psi} \left(\boldsymbol{W}_p^\top \boldsymbol{\Psi}^\top \boldsymbol{\Psi} \boldsymbol{W}_p + \boldsymbol{W}_p \boldsymbol{\Psi}^\top \boldsymbol{\Psi} \boldsymbol{W}_p^\top\right) \boldsymbol{\Phi}_i^\top
$$
$$
+ 2\lambda \boldsymbol{\Psi}_i^\top \boldsymbol{F} \boldsymbol{F}^\top \boldsymbol{\Psi} \boldsymbol{\Phi}_i^\top. \quad (19)
$$

From the Karush-Kuhn-Tucker (KKT) complementary slackness conditions, we obtain $\boldsymbol{\Theta}_i \odot \boldsymbol{H}_i = \boldsymbol{0}$, which is the fixed point equation that the solution must satisfy at convergence. By solving it, we derive the following update rule for $\boldsymbol{H}_i$:

$$
\boldsymbol{H}_i \leftarrow \boldsymbol{H}_i \odot \quad (20)
$$

$$
\left[\frac{\boldsymbol{\Psi}_i^\top (\boldsymbol{A}^\top \boldsymbol{\Psi} \boldsymbol{W}_p + \boldsymbol{A} \boldsymbol{\Psi} \boldsymbol{W}_p^\top + \lambda [\boldsymbol{F} \boldsymbol{F}^\top \boldsymbol{\Psi}]^-) \boldsymbol{\Phi}_i^\top}{\boldsymbol{\Psi}_i^\top (\boldsymbol{\Psi} \boldsymbol{W}_p^\top \boldsymbol{\Psi}^\top \boldsymbol{\Psi} \boldsymbol{W}_p + \boldsymbol{\Psi} \boldsymbol{W}_p \boldsymbol{\Psi}^\top \boldsymbol{\Psi} \boldsymbol{W}_p^\top + \lambda [\boldsymbol{F} \boldsymbol{F}^\top \boldsymbol{\Psi}]^+) \boldsymbol{\Phi}_i^\top} \right]^{\frac{1}{4}}
$$

where we separate the positive and negative parts of an arbitrary matrix $\boldsymbol{B}$ into $\boldsymbol{B}^+ = \max(\boldsymbol{B}, 0)$ and $\boldsymbol{B}^- = -\min(\boldsymbol{B}, 0)$, such that the main matrix is conveniently the addition of the negative and positive parts $\boldsymbol{B} = \boldsymbol{B}^+ - \boldsymbol{B}^-$.

IX. APPENDIX 2: ABLATION STUDIES

A. Selecting $\lambda^\star$ and Bracketing Values

We sweep $\lambda \in \{10^{-3}, 10^{-2}, \ldots, 10^3\}$ and, for each dataset (and the k used in the main results), form the utility–fairness set $\{(Q(\lambda), \bar{B}(\lambda))\}$. We retain the Pareto front (undominated points), min–max scale $(Q, \bar{B})$ to $[0, 1]$ (per dataset and k), and select

$$
\lambda^\star = \arg\min_{\lambda \in \Lambda_{\mathrm{P}}} \left\| (\tilde{Q}(\lambda), \tilde{B}(\lambda)) - (1, 1) \right\|_2,
$$

with a tie–breaker preferring smaller $|\tilde{Q} - \tilde{B}|$ (closer to the identity guide). We also report a bracket, i.e., the nearest available grid values to $\{\lambda^\star/10, 10\lambda^\star\}$, as a transparent operating band. Table VII summarizes $\lambda^\star$, the selected layer sizes, and the achieved Q and $\bar{B}$. Moreover, it reports the bracketing values λ_{lo} and λ_{hi} related to each $\lambda^\star$ and the corresponding low and high Q and $\bar{B}$ values.

a) What the table shows: (i) Datasets with strong community signal (e.g., SBM) select small $\lambda^\star$, preserving Q while still improving $\bar{B}$; The same pattern applies to NBA, and LastFM (ii) highly imbalanced or sparse social graphs (e.g., Diaries and Facebook) push $\lambda^\star$ higher to achieve parity; (iii) for medium-scale, noisy graphs (DrugNet, Friendship), $\lambda^\star$ sits near 10^{-1}, striking a balanced operating point. Brackets are tight where fronts are steep (clear trade-offs) and wider where fronts are flat (multiple near-equivalent settings). We use $\lambda^\star$ for main tables; Pareto plots in the paper illustrate the surrounding spectrum (including extreme settings within the bracket).

B. Sensitivity to the Number of Clusters (k)

Reporting & robustness. All entries are mean values over 10 random seeds (std. omitted for space); the same seeds are reused across λ for a given dataset and k. Although $\lambda^\star$ is selected using min–max–normalized $(Q, \bar{B})$ per dataset and k, a check with a simple linear scalarization $0.5\,Q + 0.5\,\bar{B}$ chose the same or bracket-adjacent setting in every case (i.e., within $\{\lambda^\star/10, 10\lambda^\star\}$). Width patterns follow fixed templates (e.g., $[64, k]$ or $[256, 64, k]$) across λ; the observed increase of $\lambda^\star$ with k persists under fixed-width variants, indicating the trend is driven by partition granularity rather than capacity. For *LastFM*, the steeper drop of $\bar{B}$ as k grows aligns with its high homophily/sparsity: finer partitions leave fewer cross-group ties to balance without larger λ. The dynamics of λ and k are illustrated in Fig 8.

Table VIII examines how the selected $\lambda^\star$ and the attained $(Q, \bar{B})$ evolve as k increases from 3 to 8. Three consistent trends emerge. (1) $\lambda^\star$ *generally rises with* k, reflecting that enforcing proportional representation becomes harder when clusters are smaller: Facebook shows a monotone increase

TABLE VII: Selected $\lambda^\star$ per dataset (with layers) and bracketing values. Q and $\bar{B}$ are raw scores at each λ; λ_{lo} and λ_{hi} are the nearest grid values to $\{\lambda^\star/10, 10\lambda^\star\}$.

Dataset	k	Layers@ $\lambda^\star$	$\lambda^\star$	λ_{lo}	λ_{hi}	$Q(\lambda^\star)$	$\bar{B}(\lambda^\star)$	$Q(\lambda_{\text{lo}})$	$\bar{B}(\lambda_{\text{lo}})$	$Q(\lambda_{\text{hi}})$	$\bar{B}(\lambda_{\text{hi}})$
NBA	2	[64, 2]	0.05	0.005	0.5	0.134	0.438	0.130	0.361	0.131	0.370
Pokec-n	2	[512, 16, 2]	10	10	10	0.164	0.177	0.164	0.177	0.164	0.177
Pokec-z	2	[512, 16, 2]	10	10	10	0.162	0.175	0.162	0.175	0.162	0.175
Diaries	5	[64, 16, 5]	50	5	500	0.716	0.787	0.745	0.688	0.517	0.802
DrugNet	5	[64, 5]	0.1	0.01	1	0.591	0.162	0.600	0.121	0.572	0.117
Facebook	5	[64, 5]	100	10	1000	0.503	0.768	0.512	0.638	0.461	0.752
Friendship	5	[64, 5]	0.1	0.01	1	0.666	0.614	0.670	0.641	0.561	0.663
LastFM	5	[256, 64, 5]	0.005	0.001	0.05	0.420	0.091	0.660	0.046	0.629	0.020
SBM	5	[64, 16, 5]	0.5	0.05	5	0.114	1.000	0.114	1.000	0.078	0.756

TABLE VIII: **Sensitivity of DFNMF to** k through the lens of $\lambda^\star$ on DrugNet, LastFM, and Facebook. Values reported at $\lambda^\star$ per k (Pareto-selected).

Dataset	k	$\lambda^\star$	Layers	Q	$\bar{B}$
DrugNet	3	0.01	[64, 3]	0.482	0.170
	5	0.10	[64, 5]	0.591	0.162
	8	0.50	[64, 8]	0.456	0.158
LastFM	3	0.005	[256, 64, 3]	0.516	0.147
	5	0.005	[256, 64, 5]	0.420	0.091
	8	0.050	[256, 64, 8]	0.384	0.090
Facebook	3	60	[64, 3]	0.408	0.818
	5	100	[64, 5]	0.503	0.768
	8	150	[64, 8]	0.437	0.652

Fig. 8: Selected $\lambda^\star$ vs. k (log-scale on y).

$(60 \to 100 \to 150)$; DrugNet increases then plateaus $(0.01 \to 0.10 \to 0.50)$; LastFM is stable at small k and rises at $k=8$ $(0.005 \to 0.005 \to 0.05)$. (2) *Modularity often peaks at a moderate* k: DrugNet and Facebook achieve their best Q at $k=5$ (0.591 and 0.503, respectively) and then drop at $k=8$, while LastFM (sparser, more homophilous) shows a steady decline $(0.516 \to 0.420 \to 0.384)$ as partitions become finer. (3) *Balance degrades as* k *grows*: the drop is mild on DrugNet $(0.170 \to 0.162 \to 0.158)$, pronounced on Facebook $(0.818 \to 0.768 \to 0.652)$, and sharp early on for LastFM $(0.147 \to 0.091 \to 0.090)$, consistent with tighter per-cluster parity constraints at smaller cluster sizes. Layer choices at $\lambda^\star$ remain compact and systematic—$[64, k]$ for DrugNet/Facebook and $[256, 64, k]$ for LastFM—supporting reproducible configurations across k.

a) Takeaway.: As k increases, maintaining per-cluster parity becomes more challenging: $\lambda^\star$ typically needs to grow,

$\bar{B}$ tends to fall, and Q often peaks around a moderate k. Reporting $\lambda^\star$ alongside $(Q, \bar{B})$ at each k provides a transparent, reproducible operating point per dataset.

C. Extended Interpretability Rsults

This appendix complements Sec. V-E2 by reporting the *full* soft-membership matrices used in the 60-node example (Fig. 5). Recall the hierarchical mapping $\boldsymbol{\Psi} = \boldsymbol{H}_1 \boldsymbol{H}_2$, where $\boldsymbol{H}_1$ encodes node→micro-cluster affinities and $\boldsymbol{H}_2$ maps micro-clusters→communities. For visual context in the main text, see the paired depiction of $\boldsymbol{H}_2$ and the trimmed $\boldsymbol{\Psi}$ in Fig. 6, and the compact slice of $\boldsymbol{H}_1$ in Table V. The full results are now illustrated in Tables IX, X, and XI

a) Reading the matrices.: Entries are nonnegative *soft memberships* (not necessarily normalized). Two patterns underpin interpretability: (i) **micro-sparsity**—rows of $\boldsymbol{H}_1$ are typically few-peaked, so most nodes participate in a small number of micro-clusters (but this highly depends on graph sructure and properties); and (ii) **structured aggregation**—columns of $\boldsymbol{H}_2$ are near one-hot for core micro-clusters (e.g., A→I, B→II, C→III), while boundary micro-clusters spread mass across communities. Consequently, $\boldsymbol{\Psi} = \boldsymbol{H}_1 \boldsymbol{H}_2$ consolidates concentrated rows into near one-hot community affinities, while mixed rows remain soft. This traceability (node→micro→community) enables transparent auditing of how local structure (under fairness regularization) aggregates into global communities.

X. Appendix 3: Intersectional Multi-Attribute Fairness

A. Constructing the Intersectional Fairness Matrix

The problem of fairness often gets amplified for individuals belonging to the intersection of two or more protected/sensitive attributes (e.g. 'Race', and 'Gender') [45]. Suppose each node has A sensitive attributes. Attribute $a \in \{1, \ldots, A\}$ has m_a groups and one-hot indicator $G^{(a)} \in \mathbb{R}^{n \times m_a}$. To enforce *intersectional* (joint) parity across the Cartesian product of all attributes as previously done in [46], [47], we build a joint one-hot matrix:

$$G_{\text{int}} \in \mathbb{R}^{n \times M}, \qquad M = \prod_{a=1}^{A} m_a, \qquad (21)$$

whose columns correspond to all joint categories (e.g., *male∧Asian*, *female∧White*, …). Concretely, each joint column

TABLE IX: Full node→micro-cluster soft memberships (H_1, micro-clusters A–L).

Node	A	B	C	D	E	F	G	H	I	J	K	L
1	0.26	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.01	0.00	0.00	0.00
2	0.25	0.00	0.01	0.00	0.01	0.09	0.02	0.00	0.01	0.00	0.04	0.00
3	0.21	0.00	0.01	0.00	0.09	0.03	0.00	0.00	0.00	0.00	0.11	0.00
4	0.23	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00
5	0.31	0.00	0.02	0.00	0.01	0.15	0.05	0.10	0.05	0.02	0.09	0.00
6	0.05	0.00	0.00	0.00	0.00	0.26	0.61	0.00	0.00	0.00	0.00	0.00
7	0.03	0.00	0.00	0.00	0.00	0.17	0.81	0.00	0.00	0.00	0.00	0.00
8	0.01	0.00	0.00	0.00	0.00	0.04	0.81	0.00	0.00	0.00	0.00	0.00
9	0.19	0.00	0.00	0.00	0.00	0.33	0.23	0.00	0.01	0.00	0.00	0.00
10	0.11	0.00	0.03	0.00	0.00	0.16	0.52	0.00	0.00	0.00	0.07	0.00
11	0.10	0.02	0.00	0.00	0.00	0.05	0.00	0.50	0.03	0.00	0.00	0.00
12	0.06	0.00	0.00	0.00	0.00	0.03	0.00	0.63	0.08	0.00	0.00	0.00
13	0.16	0.10	0.03	0.00	0.00	0.06	0.00	0.54	0.23	0.04	0.15	0.03
14	0.07	0.06	0.00	0.00	0.00	0.03	0.00	0.56	0.05	0.00	0.00	0.00
15	0.14	0.04	0.00	0.00	0.02	0.18	0.07	0.64	0.05	0.01	0.00	0.01
16	0.13	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.03	0.00	0.00	0.00
17	0.06	0.00	0.00	0.00	0.03	0.10	0.00	0.04	0.00	0.00	0.00	0.07
18	0.19	0.00	0.00	0.00	0.09	0.01	0.00	0.01	0.00	0.00	0.00	0.00
19	0.20	0.00	0.00	0.00	0.22	0.02	0.00	0.02	0.00	0.00	0.00	0.00
20	0.22	0.00	0.03	0.00	0.02	0.04	0.00	0.00	0.11	0.00	0.06	0.65
21	0.00	0.06	0.00	0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.31	0.00
22	0.03	0.00	0.00	0.28	0.00	0.05	0.00	0.00	0.01	0.50	0.00	0.05
23	0.00	0.00	0.00	0.31	0.00	0.00	0.00	0.00	0.00	0.57	0.00	0.00
24	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.57	0.00	0.00
25	0.00	0.00	0.01	0.40	0.04	0.00	0.00	0.00	0.00	0.86	0.00	0.00
26	0.00	0.22	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
27	0.00	0.21	0.00	0.05	0.00	0.02	0.00	0.08	0.01	0.00	0.00	0.00
28	0.00	0.23	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
29	0.00	0.28	0.01	0.02	0.06	0.00	0.00	0.02	0.00	0.00	0.12	0.00
30	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
31	0.00	0.23	0.00	0.00	0.00	0.00	0.00	0.04	0.04	0.00	0.00	0.00
32	0.00	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
33	0.00	0.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
34	0.00	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
35	0.00	0.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
36	0.00	0.10	0.00	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
37	0.00	0.00	0.00	0.84	0.01	0.00	0.10	0.00	0.00	0.07	0.01	0.00
38	0.00	0.01	0.54	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
39	0.00	0.05	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
40	0.00	0.09	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
41	0.00	0.00	0.00	0.18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
42	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
43	0.01	0.00	0.24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.24
44	0.00	0.00	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
45	0.00	0.00	0.26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
46	0.03	0.01	0.01	0.02	0.53	0.01	0.00	0.00	0.20	0.01	0.03	0.07
47	0.00	0.05	0.07	0.01	0.23	0.00	0.00	0.00	0.00	0.00	0.08	0.00
48	0.00	0.00	0.06	0.17	0.58	0.00	0.00	0.00	0.01	0.10	0.04	0.00
49	0.03	0.01	0.10	0.02	0.63	0.02	0.00	0.00	0.07	0.00	0.10	0.00
50	0.04	0.00	0.16	0.00	0.15	0.05	0.00	0.00	0.02	0.00	0.11	0.00
51	0.01	0.00	0.00	0.00	0.64	0.17	0.10	0.21	0.05	0.00	0.00	0.35
52	0.00	0.00	0.00	0.00	0.27	0.00	0.02	0.03	0.00	0.00	0.00	0.40
53	0.00	0.00	0.05	0.00	0.25	0.00	0.01	0.01	0.00	0.00	0.00	0.34
54	0.00	0.00	0.00	0.00	0.11	0.01	0.02	0.07	0.01	0.00	0.00	0.06
55	0.03	0.00	0.00	0.00	0.32	0.01	0.01	0.00	0.00	0.00	0.01	0.31
56	0.01	0.00	0.24	0.00	0.04	0.13	0.00	0.01	0.00	0.00	0.25	0.00
57	0.06	0.00	0.26	0.00	0.05	0.01	0.09	0.06	0.01	0.00	0.26	0.00
58	0.00	0.00	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.24	0.00
59	0.00	0.00	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00
60	0.00	0.00	0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00

TABLE X: Micro-cluster membership values of H_2 across the three main clusters (I–III).

Micro cluster	Clusters		
	I	II	III
A	1.04	0.00	0.00
B	0.00	1.00	0.00
C	0.00	0.00	1.00
D	0.00	0.05	0.01
E	0.02	0.00	0.19
F	0.00	0.00	0.00
G	0.08	0.00	0.00
H	0.06	0.08	0.00
I	0.00	0.00	0.00
J	0.00	0.02	0.06
K	0.00	0.00	0.01
L	0.02	0.00	0.01

is the elementwise AND (Hadamard product) of one columns taken from each $G^{(a)}$:

$$G_{\text{int}} = G^{(1)} \odot G^{(2)} \odot \cdots \odot G^{(A)}. \qquad (22)$$

We then apply proportional centering and drop one redundant column to avoid linear dependence:

$$F_{\text{int}} = G_{\text{int}} - \tfrac{1}{n}\mathbf{1}_n\big(\mathbf{1}_n^\top G_{\text{int}}\big), \qquad F_{\text{int}} \in \mathbb{R}^{n \times (M-1)}. \quad (23)$$

DFNMF's fairness penalty becomes

$$\mathcal{R}_{\text{int}}(H) = \big\|F_{\text{int}}^\top H\big\|_F^2, \qquad (24)$$

which enforces demographic parity *jointly* over all intersections. Complexity-wise, this adds $O(nk\,M)$ per iteration (for forming $F_{\text{int}}^\top H$). In our SBM study with two attributes (gender: $m_1{=}2$, ethnicity: $m_2{=}5$), $M{=}10$ is modest.

a) Tiny schematic (two attributes).: For gender $\{M, F\}$ and ethnicity $\{A, W, B, C, D\}$, the $M{=}10$ joint groups are

$$\{(M, A), (M, W), (M, B), (M, C), (M, D),$$
$$(F, A), (F, W), (F, B), (F, C), (F, D)\}.$$

Each column of G_{int} is the indicator of one joint group; F_{int} is its centered version (with one dropped column).

B. Metrics under Intersectional Fairness

We evaluate standard utility and fairness metrics:

- **Modularity** Q (higher is better).
- **Intersectional balance** $\bar{B}_{\text{int}}$ computed over the M joint groups: $\bar{B}_{\text{int}} = \frac{1}{k}\sum_{l=1}^{k}\min_{g \neq g'} \frac{|V_g \cap C_l|}{|V_{g'} \cap C_l|}$, where g, g' range over joint categories.

C. SBM at Scale: Enforcing Intersectional Parity

We compare *single-attribute* fairness (gender-only; ethnicity-only) against *intersectional* fairness (gender×ethnicity) on SBMs with $k{=}10$ clusters and $n \in \{2K, 5K, 10K\}$ nodes. We fix $\lambda{=}100$ to isolate the effect of the constraint. For single-attribute runs, $\bar{B}$ is computed on that attribute; for the intersectional run, $\bar{B}_{\text{int}}$ is computed on the $M{=}10$ joint groups. The results are illustrated in Table XII.

TABLE XI: Full node→community soft memberships ($\Psi = H_1 H_2$, communities I–III).

Node	I	II	III
1	0.27	0.00	0.00
2	0.26	0.00	0.01
3	0.22	0.00	0.03
4	0.24	0.00	0.00
5	0.33	0.01	0.02
6	0.10	0.00	0.00
7	0.10	0.00	0.00
8	0.08	0.00	0.00
9	0.22	0.00	0.00
10	0.16	0.00	0.03
11	0.13	0.06	0.00
12	0.10	0.05	0.00
13	0.20	0.14	0.03
14	0.11	0.10	0.00
15	0.19	0.09	0.00
16	0.14	0.00	0.00
17	0.07	0.00	0.00
18	0.20	0.00	0.00
19	0.21	0.00	0.00
20	0.24	0.00	0.04
21	0.00	0.08	0.02
22	0.03	0.02	0.03
23	0.00	0.03	0.04
24	0.00	0.01	0.03
25	0.00	0.04	0.07
26	0.00	0.22	0.00
27	0.00	0.22	0.00
28	0.00	0.23	0.00
29	0.00	0.28	0.02
30	0.00	0.20	0.00
31	0.00	0.23	0.00
32	0.00	0.30	0.00
33	0.00	0.26	0.00
34	0.00	0.23	0.00
35	0.00	0.26	0.00
36	0.00	0.11	0.00
37	0.01	0.04	0.01
38	0.00	0.04	0.01
39	0.00	0.05	0.00
40	0.00	0.09	0.00
41	0.00	0.00	0.18
42	0.00	0.00	0.33
43	0.02	0.00	0.24
44	0.00	0.00	0.30
45	0.00	0.00	0.26
46	0.04	0.01	0.11
47	0.00	0.05	0.11
48	0.01	0.01	0.18
49	0.04	0.01	0.22
50	0.04	0.00	0.19
51	0.05	0.02	0.13
52	0.02	0.00	0.06
53	0.01	0.00	0.10
54	0.01	0.01	0.02
55	0.04	0.00	0.06
56	0.02	0.00	0.24
57	0.07	0.01	0.26
58	0.00	0.00	0.22
59	0.00	0.00	0.31
60	0.00	0.00	0.19

TABLE XII: **SBM, intersectional multi-attribute fairness** ($k=10$, $\lambda=100$). Intersectional uses $M=10$ joint groups (gender×ethnicity).

Attribute	n	M (#groups)	Q	$\bar{B}$
Gender only	2K	2	1.000	1.000
	5K	2	1.000	1.000
	10K	2	1.000	1.000
Ethnicity only	2K	5	0.119	0.8985
	5K	5	0.124	0.9999
	10K	5	0.124	1.0000
Intersectional (G×E)	2K	10	0.110	0.3522
	5K	10	0.112	0.4263
	10K	10	0.115	0.4507

a) Discussion:

- **Single-attribute parity**. Gender-only aligns with the planted structure in this SBM (both Q and $\bar{B}$ near 1). Ethnicity-only parity achieves very high balance with modest $Q\approx0.12$, improving or stabilizing as n grows.
- **Intersectional parity**. Enforcing joint parity across $M=10$ groups is stricter: $\bar{B}_{\text{int}}$ is lower than single-attribute balances at small n, but *increases with scale* (from 0.35 at 2K to 0.45 at 10K). The utility cost is controlled (slight Q increase with n from 0.110 to 0.115), indicating that larger graphs make intersectional constraints more feasible.
- **Takeaway**. DFNMF can efficiently enforce intersectional demographic parity directly via F_{int}; The soft balanced fairness encoding introduced in Section IV-A allows this extension efficiently. The expected trade-off (tighter fairness $\Rightarrow$ lower Q) is observed, while scaling the graph improves joint feasibility without tuning λ.