D1.3 Report on Models for Bias

Christos Gkartzios[1], **Glykeria Toulina**[1], **Spyridon Tzimas**[1], **Athanasios Konstadinidis**[1], **Panagiotis Papadakos**[2], **Panagiotis Papapetrou**[3], **Evaggelia Pitoura**[1], **Panayiotis Tsaparas**[1]

[1] Department of Computer Science and Engineering, University of Ioannina, Greece
[2]Institute of Computer Science (ICS) - Foundation for Research and Technology - Hellas (FORTH), Greece
[3]Department of Computer and Systems Sciences, Stockholm University, Sweden

1 Introduction

The objective of the THEMIS project is to study bias and fairness in AI algorithms and machine learning pipelines. In Deliverables D1.1 and D1.2, we surveyed existing approaches and metrics for defining and measuring bias in tasks such as classification, clustering, community detection, and network analysis, and we described in detail the metrics relevant to our research. In this deliverable, we present the novel metrics and models for bias and fairness developed within the project.

Most of the models and metrics considered in this deliverable fall under the category of *group fairness*. Specifically, we assume that data instances are partitioned into groups based on the value of a sensitive attribute, such as gender, race, or religion. Group fairness requires that these groups are treated equitably by the algorithm. In some cases, rather than enforcing equal treatment across all groups, we assume the existence of a *protected group* for which preferential treatment is desirable. Such a group may correspond to an underrepresented minority that the algorithm aims to support. For example, in a hiring scenario, women may constitute the protected group, with the goal of increasing their representation among the selected candidates. Most of the fairness notions we consider fall under the broader category of *representation fairness*, which requires that protected groups are adequately represented in the algorithm's output.

The contributions of this deliverable span the following directions:

- **Classification and Counterfactual Generation:** We propose a novel measure of classification bias that relies on the burden of counterfactual explanations generation, as well as a measure of fairness for counterfactual generation.
- **Community Detection**: We introduce new metrics for community fairness based on group connectivity and modularity.
- **Opinion Formation:** We propose a novel notion of fairness for opinion formation processes based on influence.
- **k-core:** We propose a novel measure of fairness for the graph k-core.
- **Network Homophily:** We conduct an LLM-agent–based simulation to understand and model the emergence of homophily in social networks.

The rest of the report is structured as follows. In Section 2 we present our fairness metric based on counterfactual burden. In Section 3 we present our novel fairness metrics for community detection. In Section 4 we present a novel metric for opinion formation fairness. In Section 5 we present metrics for measuring the fairness of the graph k-core. In Section 6 we present a simulation for modeling network homophily. Finally, Section 7 concludes the paper.

2 Counterfactual Generation and Classification bias

In this section, we present a novel measure for classification bias, via the use of counterfactual explanations, and a methodology for introducing fairness in counterfactual explanations. This work was published in ICDM 2024 [19]. The paper was selected as one of the best-ranked papers of ICDM 2024, and an extended version of the work was published in Knowledge and Information Systems (KAIS) [21].

Counterfactual explanations help us understand opaque machine learning (ML) models by exploring 'what-if' scenarios for individual instances [13]. The word *counterfactual* may be used as a noun to refer to the instances that are counterfactual points themselves, or as an adjective to describe the points, the explanations, or the reasoning itself. We will use the shorthand *CF* to abbreviate the word "counterfactual". Given a dataset and a trained classifier that maps input instances to class labels, CF explanations can highlight the relevant feature value changes for an instance of interest that would result in an alternative predicted class label [16, 26, 13]. Consequently, a CF is also known as a *recourse* [15], since it suggests actions to improve the situation of a given instance [31, 13, 16, 15, 20]. For example, CF explanations may highlight the changes on the features of an individual (e.g., marital status, habits, education, occupation) to obtain a positive answer on a loan application, or to move from a low-wealth to a high-wealth status [31, 15]. Usually, the closest point with the desired label is selected as a CF for the given instance [13], since that reduces the feature changes the instance must apply to reach the desired label.

Nevertheless, with sensitive features, such as gender, race, or age, the suggested changes may hide biases towards the sensitive groups, which are not easy to detect or measure. These biases, if left undetected or unattended, could lead to unfair and harmful outcomes. The assessment of model biases or algorithmic fairness through the recommendations suggested by CFs is known as *counterfactual fairness* [20, 17, 2, 22]. For example, consider the fairness assessment of two commonly used public datasets: Adult[1] and COMPAS[2]. For Adult, the class label indicates whether a person earns more than \$50K/year or not, while for COMPAS it indicates whether a person is a recidivist (a person recommitting crimes) or not. Fig. 1 illustrates the average difficulty in achieving the desired state (i.e., high wealth or no recidivism) in the form of a measure called *burden* [31, 15]. Burden is the distance between an instance and its closest CF, and the figure shows its aggregated value per sensitive group. We observe a higher aggregated burden for females than for males in wealth prediction (Fig. 1a), implying that it is harder for females to achieve higher wealth. Moreover, we observe that it is harder for males than for females to not be recidivists (Fig. 1b). Similarly, it is harder for non-white people to achieve higher wealth, and harder for African-Americans to not be recidivists.

Formally, let $\mathcal{X}$ be a heterogeneous feature space with binary, categorical, ordinal, and continuous features. A dataset $\mathcal{D}$ is a collection of n pairs of (X, y) where X is a data sample (i.e., instantiation) of $\mathcal{X}$ and y is its corresponding binary class label $y \in \{\text{“}-\text{”}, \text{“}+\text{”}\}$. $\mathcal{D}$ is divided into a training and a test set, denoted as $\mathcal{D}_{Train}$ and $\mathcal{D}_{Test}$, respectively.

Moreover, let $\mathcal{S} \subseteq \mathcal{X}$ be a set of sensitive features in $\mathcal{X}$, such as *sex* or *race*. Each sensitive feature $s \in \mathcal{S}$ may be used to define different *sensitive groups* of data samples. A sensitive group of feature s is denoted as s_k, where k defines a condition on that feature, which is denoted by function $cond(\cdot)$. If s satisfies condition k then $cond(s,k) == \text{'true'}$. For example, sensitive feature *sex* may be used to define two sensitive groups, i.e., s_{female} and s_{male}, corresponding to data samples for which *sex*==‘female’ and *sex*==‘male’, respectively. Given a classifier $f(\cdot)$, we define the set of false-negative test instances in sensitive group s_k as:

$$\mathcal{D}^{s_k}_{TestFN} = \{(X, \text{“}+\text{”}) | f(X) = \text{“}-\text{”}, cond(s,k), X \in \mathcal{D}_{Test}\}$$

For each instance $X_i \in \mathcal{D}^{s_k}_{TestFN}$ we use a CF generator to get its CF, X'_i. Let $d(X_i, X'_i)$ denote the distance between X_i and X'_i, defined as a combination of the $L1$-norm and the $L0$-norm. This is the burden incurred by instance X_i in trying to attain the feature values of X'_i. The *Accuracy*

[1] https://archive.ics.uci.edu/dataset/2/adult
[2] https://www.kaggle.com/datasets/danofer/compass

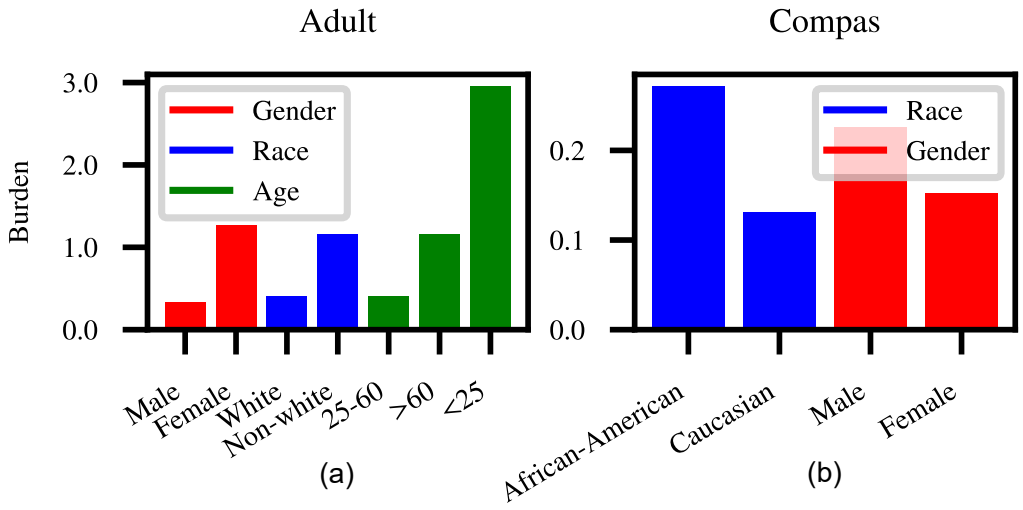

Figure 1: Burden for sensitive groups (x-axis) belonging to different features (colors), showing biases on Adult and COMPAS.

Weighted Burden(AWB) measure, introduced in [20], is the product of predictive equality (the false negative ratio) and the average burden per sensitive group. It is calculated as follows:

$$\text{AWB}^{s_k} = \frac{\sum\limits_{X_i \in \mathcal{D}^{s_k}_{TestFN}} d(X_i, X_i')}{|\{(X, y) \in \mathcal{D}_{Test} | cond(s, k), y = \text{“}+\text{”}\}|}, \tag{1}$$

where the denominator is the amount of true positives in the sensitive group s_k. Eq. 1 indicates that a higher number of false negatives, or a higher distance between each instance and its CF in the s_k sensitive group, make the AWB^{s_k} burden higher.

To measure bias, we define the cost (C_{fair}) in generating CFs, associated with the presence of biases, by estimating the differences in burden between the sensitive groups. We define $\text{AWB}_{min} = \min_{s_k} \text{AWB}^{s_k}$ and $\text{AWB}_{max} = \max_{s_k} \text{AWB}^{s_k}$ as the minimum and maximum burden, respectively, over all sensitive groups. C_{fair} is then defined as the absolute difference between these two terms:

$$C_{fair} = \text{AWB}_{max} - \text{AWB}_{min} \tag{2}$$

A high value of C_{fair} indicates a strong discrepancy in the CF burden for the different sensitive groups. That is, there is a group for which, on average, there are cheap recourse actions for reversing an undesirable decision, and a group for which, on average, recourse actions are expensive, indicating unfairness and bias of the classifier in the treatment of the different groups.

This definition of bias and unfairness is simple and intuitive and lends itself to algorithmic treatment for obtaining fair CFs. Therefore, the same measure can also be used to assess the fairness of a collection of counterfactual explanations.

3 Community Detection Fairness

Networks capture relationships between entities across diverse domains, including social platforms, scientific collaboration, and citation systems. In many such networks, nodes tend to form communities, i.e., subsets of nodes that exhibit higher internal connectivity relative to the rest of the network [23, 7]. These communities play a critical role in determining how information spreads and how opinions are shaped [35, 6].

Traditional community detection algorithms aim to maximize quality, typically optimizing metrics that capture the intra-community connectivity compared to the inter-community connectivity. Modularity is a commonly used such metric. However, such algorithms often neglect fairness considerations. In many real-world networks, nodes carry sensitive attributes such as gender, age, or ethnicity, which naturally partition the network into groups. Recent research in network algorithmic fairness has emphasized the importance of equitable treatment, particularly at the group level [25, 29, 30]. In this work, we focus on the fairness of community detection algorithms on networks.

Most previous research on group fairness of communities asks that the representation of groups within each community is balanced [4, 3, 18]. The balance fairness metric was described in detail in Deliverables D1.1 and D1.2, and we consider algorithms for discovering balanced communities and clusters (see Deliverable D3.2).

In this project, we introduce a novel fairness metric for communities that shifts the focus from nodes to connections. We ask the key question, whether each group is equally well-connected within each community. For example, consider a collaboration network. Do women in the network participate in an equitable number of connections within the formed communities? The strength of connections within each community is vital for minorities to be heard, and influence others. Our work has been resulted in two conference publications [10, 12], and an upcoming submission to the KAIS journal.

3.1 Modularity-based Community Detection

Given as input a network $G = (V, E)$, the output of a community detection algorithm is a partition of the nodes into k disjoint subsets (communities), $\mathcal{C} = \{C_1, C_2, ..C_k\}$, $C_i \subseteq V$, $C_i \cap C_j = \emptyset$, $\cup_{i=1}^{k} C_i = V$. The number of communities k may be given as input, or it may be decided by the algorithm. The goal is to discover communities where the nodes are densely intra-connected, while sparsely inter-connected.

Modularity measures the divergence between the number of intra-community edges and the expected such number assuming a null model [27, 5]. The most commonly used null model is a random graph where the expected degree of each node within the graph is equal to the actual degree of the corresponding node in the real network. Specifically, the modularity of community $\mathcal{C}_i$, $Q(\mathcal{C}_i)$, is defined as [27]:

$$Q(C_i) = \frac{1}{2m}\left(\sum_{u \in C_i}\sum_{v \in C_i} A_{uv} - \frac{k_u\, k_v}{2m}\right) \tag{3}$$

where A is the adjacency matrix of G, m the number of edges in G and k_u, k_v the degree of node u, and v respectively. Modularity provides a measure of how well nodes in a community are connected with each other. Negative values indicate less connections than expected, while positive values indicate more connections.

To define fairness, we assume that the nodes of the graph are associated with some sensitive attribute A, such as gender, religion or race, that takes t values $\{a_1, ..., a_t\}$, which partition the nodes of the graph into t *groups* $G = \{G_1, G_2, .., G_t\}$, $G_i = \{v \in V : A(v) = a_i\}$. In the following, we will often refer to the attribute values, and the corresponding groups, as *colors*. For simplicity, we assume two colors Red (R) and Blue (B). We will assume that the red group is the *protected* or minority group, for which we want to mitigate bias.

3.2 Group Modularity fairness

Our goal is to ensure that red nodes are well connected within each community. Thus, for each red node u in C_i we take the difference between the actual number of its intra-community edges and the expected such number. We call this measure *red modularity*.

As before, the expected number of connections is estimated assuming as null model a random graph that preserves the degrees of nodes in G. Using this null model, red modularity, $Q^R(C_i)$ is

defined as:

$$Q^R(C_i) = \frac{1}{2m} \sum_{u \in C_i^R} \sum_{v \in C_i} \left(A_{uv} - \frac{k_u\, k_v}{2m} \right). \tag{4}$$

We define similarly the *blue modularity* $Q^B(C_i)$. We refer to red and blue modularity collectively as *group modularity*.

Note that if we consider the whole graph as a single community both the red and the blue modularity are zero. In general, positive values in a community mean that the nodes with the corresponding color are more connected in the community than expected.

We define *(group) modularity unfairness* by comparing the red and blue modularity.

Definition 1. *For a community $C_i \in \mathcal{C}$, the modularity unfairness of C_i, $u(C_i)$, is defined as:*

$$u(C_i) = Q^R(C_i) - Q^B(C_i).$$

Negative values of $u(C_i)$ indicate unfairness towards the red group meaning that the red nodes are less well-connected within the community than the blue ones. Positive values indicate the opposite, while a zero value indicates lack of unfairness towards any of the groups.

We also consider diversity within each community by looking at the edges that connect nodes of different color. Let us call these edges *diverse edges*. Note that the expected number of diverse edges cannot be estimated using the same null model, since we need to know the color of both endpoints of each edge. Instead, in this case, we estimate the expected number of diverse edges using as null model a random bipartite graph, with edges only between nodes of different color, that preserves the degrees of the nodes in the original graph G.

For a community C_i, the *diversity modularity*, or simply *diversity*, is defined as:

$$D^{RB}(C_i) = \frac{1}{2m} \sum_{u \in C_i^R} \sum_{v \in C_i^B} \left(A_{uv} - \frac{k_u\, k_v}{m} \right). \tag{5}$$

If we consider the whole graph as a single community, then diversity takes a non positive value. The larger the value of D^{RB} the more diverse the network.

We also consider a null model which is not agnostic of the color of edge endpoints. For a node u, let k_u^R be the number of edges of u to red nodes and k_u^B be the number of edges of u to blue nodes, $k_u^R + k_u^B = k_u$. In the following, k_u^R and k_u^B are respectively called the *red degree* and *blue degree* of node u.

We consider as null model a random graph where the expected red degree and the expected blue degree of each node is equal to the actual red degree and blue degree of the corresponding node in the real graph G. Formally, let P_{uv} be the probability of creating an edge between nodes u and v. Let m_{RR} be the number of red-red edges, m_{RB} the number of red-blue edges and m_{BB} the number of blue-blue edges in the graph. We have that $P_{uv} = k_u^R\, k_v^R / 2m_{RR}$, for red nodes $u, v \in R$, $P_{uv} = k_u^B\, k_v^B / 2m_{BB}$ for blue nodes $u, v \in B$, and $P_{uv} = k_u^B\, k_v^R / m_{RB}$ for red-blue nodes $u \in R$ and $v \in B$. For any node u, it holds that $\sum_{v \in R} P_{uv} = k_u^R$ and $\sum_{v \in B} P_{uv} = k_u^B$.

We define the *labeled red modularity* $Q_L^R(C_i)$ by taking again the difference between the actual number of intra-community edges involving red nodes, and the expected such number, but now considering the color (or, in general, label) of both endpoints.

$$\begin{aligned} Q_L^R(C_i) = \frac{1}{2m} (& \sum_{u \in C_i^R} \sum_{v \in C_i^B} (A_{uv} - \frac{k_u^B\, k_v^R}{m_{RB}}) \\ & + \sum_{u \in C_i^R} \sum_{v \in C_i^R} (A_{uv} - \frac{k_u^R\, k_v^R}{2m_{RR}})). \end{aligned} \tag{6}$$

We define similarly the *labeled blue modularity* $Q_L^B(C_i)$. We refer to labeled red and labeled blue modularity collectively as *labeled group modularity*. Again, if we consider the whole graph as a single community both the labeled red and the labeled blue modularity are zero.

We define the *labeled modularity unfairness* by comparing the red and blue labeled modularity.

Definition 2. *For a community $C_i \in \mathcal{C}$, the labeled modularity unfairness of C_i, $u_L(C_i)$, is defined as:*

$$u_L(C_i) = Q_L^R(C_i) - Q_L^B(C_i).$$

Negative values of $u_L(C_i)$ indicate unfairness towards the red group, positive values indicate unfairness towards the blue group, and a zero value lack of unfairness.

We define *labeled diversity modularity*, or simply *labeled diversity*, as follows:

$$D_L^{RB}(C_i) = \frac{1}{2m}\left(\sum_{u \in C_i^R}\sum_{v \in C_i^B}\left(A_{uv} - \frac{k_u^B\, k_v^R}{m_{RB}}\right)\right). \tag{7}$$

The labeled diversity of the whole graph is zero, while positive diversity values in a community indicate that the community contains more diverse edges than expected.

3.3 Group-Aware Modularity Matrices

Building upon the work in [10], we now consider a similar definition of modularity fairness, and we propose modifications of the Spectral and Deep Community Detection algorithms that incorporate fairness.

Let $A \in \mathbb{R}^{n \times n}$ denote the adjacency matrix of the graph G. We partition A into four disjoint sub-matrices:

$$A = \begin{bmatrix} A_{RR} & A_{RB} \\ A_{BR} & A_{BB} \end{bmatrix}$$

where

- A_{RR}: the matrix with the edges between Red nodes,
- A_{RB}: the matrix with the edges from Red nodes to Blue nodes,
- A_{BR}: the matrix with the edges from Blues nodes to Red nodes, and
- A_{BB}: the matrix with the edges between Blue nodes.

Since the graph is undirected, it holds that $A_{RB} = A_{BR}^\top$.

We now define the sub-matrices A_R, A_B, and A_{div} as follows:

$$A_R = \begin{bmatrix} A_{RR} & A_{RB} \\ A_{BR} & \mathbf{0} \end{bmatrix} \quad A_B = \begin{bmatrix} \mathbf{0} & A_{RB} \\ A_{BR} & A_{BB} \end{bmatrix} \quad A_{\text{div}} = \begin{bmatrix} \mathbf{0} & A_{RB} \\ A_{BR} & \mathbf{0} \end{bmatrix}$$

where

- A_R: the matrix with all edges incident to Red nodes,
- A_B: the matrix with all edges incident to Blue nodes
- A_{div}: the inter-group adjacency matrix, capturing diversity across groups.

Given these matrices, we can define the corresponding subgraphs G_R, G_B and G_{div}. We will use these matrices to decompose the modularity matrix, and define the clustering objectives that we use throughout this work.

Classical modularity optimization builds upon the *modularity matrix*, introduced by Newman [28]:

$$B = A - \frac{dd^\top}{2m}$$

where A is the adjacency matrix, d is the degree vector with d_i being the degree of node i, and m is the number of edges in the graph.

The modularity score for the partition $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ can be computed using the modularity matrix B. Let $S \in \{0,1\}^{n \times k}$ be the binary community assignment matrix, where $S_{ij} = 1$ if node $i \in C_j$, and 0 otherwise. Then, the modularity of the partition $\mathcal{C}$ defined by the assignment matrix S is given by:

$$Q(\mathcal{S}) = \frac{1}{2m} \operatorname{Tr}(S^\top B S), \tag{8}$$

where $\operatorname{Tr}(\cdot)$ denotes the matrix trace [28].

We can now use the decomposition of the adjacency matrix to define the group-aware variants of modularity defined in [10], by decomposing the modularity matrix. Specifically, we define the *red modularity matrix* B_R using the red adjacency matrix A_R as follows:

$$B_R = A_R - \frac{d_R d_R^\top}{2m_R}$$

where d_R is the degree vector for the graph G_R and m_R is the number of edges in the graph G_R. The modularity score for red group connectivity, denoted Q_{R}, is then given by:

$$Q_{\text{R}}(S) = \frac{1}{2m} \operatorname{Tr}(S^\top B_R S) \tag{9}$$

Analogously, we define the *blue modularity matrix* B_B using the blue adjacency matrix A_B and the degree vector d_B of the graph G_B. We also define the *diversity modularity matrix* B_{div} using the matrix A_{div} and the corresponding degree vector d_{div}:

$$B_{\text{div}} = A_{\text{div}} - \frac{d_{\text{div}} d_{\text{div}}^\top}{2m_{\text{div}}}$$

$$Q_{\text{div}}(S) = \frac{1}{2m} \operatorname{Tr}(S^\top B_{\text{div}} S) \tag{10}$$

These formulations allow us to measure and optimize modularity with respect to both group-specific and inter-group connectivity patterns. By incorporating group constraints directly into the adjacency structure, our modularity matrix variants enable fairness-aware spectral optimization, and fairness-aware loss functions.

3.4 Multi-Group Fairness in Loss-Based Deep Models

So far, we focused on a binary sensitive attribute with two groups, red and blue, and corresponding group-modularity scores $Q_{\text{R}}(S)$ and $Q_B(S)$. We now extend the loss-based deep models to the multi-group setting, where nodes may belong to more than two groups.

Let $a : V \to \mathcal{U}$ be a sensitive attribute with $|\mathcal{U}| \geq 2$, and let $V_g = \{u \in V : a_u = g\}$ denote the set of nodes in group $g \in \mathcal{U}$.

3.4.1 Max-Min Group Modularity

Definition 3 (Multi-group group modularity)**.** *For each group $g \in \mathcal{U}$, we define the group adjacency matrix A_g^{grp} by keeping all edges incident to group g:*

$$(A_g^{grp})_{uv} = \begin{cases} A_{uv}, & \text{if } a_u = g \text{ or } a_v = g, \\ 0, & \text{otherwise.} \end{cases}$$

Let d_g^{grp} be the degree vector of A_g^{grp} and let m_g^{grp} be the number of edges in the corresponding subgraph. We define the group modularity matrix

$$B_g^{grp} = A_g^{grp} - \frac{d_g^{grp}(d_g^{grp})^\top}{2m_g^{grp}},$$

and the group modularity score of g under assignment S as

$$Q_g(S) = \frac{1}{2m} \operatorname{Tr}(S^\top B_g^{grp} S).$$

In the binary case $\mathcal{U} = \{R, B\}$, this recovers $Q_{\mathrm{R}}(S)$ and $Q_B(S)$.

To measure group connectivity from a worst-case perspective, we define the minimum group modularity:

Definition 4 (Max-min group modularity)**.** *The max-min group modularity is defined as follows:*

$$Q_{\min}^{grp}(S) = \min_{g \in \mathcal{U}} Q_g(S).$$

Maximizing $Q_{\min}^{\mathrm{grp}}(S)$ encourages partitions where even the least well connected group achieves strong within-community connectivity.

Multi Deep Group Modularity: We extend DeepGroupby targeting the worst connected group, rather than selecting a single target group. The loss function is:

$$\mathcal{L}_{\text{Multi Deep Group Modularity}} = -Q(S) \; - \; \lambda\, Q_{\min}^{\mathrm{grp}}(S) \; + \; \gamma\, \mathcal{R}_{\mathrm{collapse}},$$

where λ controls the tradeoff between overall modularity and the max–min group objective.

3.4.2 Max-Min Group Diversity

We next define a multi-group diversity objective without using pairwise group pairs. Instead, we measure, for each group g, how well it connects to the rest of the population inside the discovered communities.

Definition 5 (Group-to-rest diversity modularity)**.** *For each group $g \in \mathcal{U}$, we define a group-to-rest diversity adjacency matrix A_g^{div} by keeping only edges with exactly one endpoint in g:*

$$(A_g^{div})_{uv} = \begin{cases} A_{uv}, & \text{if } (a_u = g \wedge a_v \neq g) \text{ or } (a_v = g \wedge a_u \neq g), \\ 0, & \text{otherwise.} \end{cases}$$

Let d_g^{div} be the degree vector of A_g^{div} and let m_g^{div} be the number of edges in the corresponding subgraph. We define

$$B_g^{div} = A_g^{div} - \frac{d_g^{div}(d_g^{div})^\top}{2m_g^{div}}, \qquad Q_g^{div}(S) = \frac{1}{2m} \operatorname{Tr}(S^\top B_g^{div} S).$$

In the binary case, $A_R^{div} = A_B^{div} = A_{div}$ and therefore $Q_g^{div}(S)$ coincides with $Q_{\mathrm{div}}(S)$.

Definition 6 (Max-min group diversity)**.** *The max-min group diversity is defined as follows:*

$$Q_{\min}^{div}(S) = \min_{g \in \mathcal{U}} Q_g^{div}(S).$$

Maximizing $Q_{\min}^{div}(S)$ encourages partitions where every group has strong cross-group connectivity within the communities.

Multi Deep Group Diversity: We extend DeepDiversityto the multi-group setting by maximizing the minimum group-to-rest diversity modularity. The loss function is:

$$\mathcal{L}_{\text{Multi Deep Group Diversity}} = -Q(S) \; - \; \lambda\, Q_{\min}^{div}(S) \; + \; \gamma\, \mathcal{R}_{\text{collapse}},$$

where λ controls the tradeoff between modularity and diversity.

3.4.3 Multi-Group Fairness

Finally, we generalize the modularity-gap notion of unfairness to more than two groups. A partition is fair when group modularities are close to each other.

Definition 7 (Multi-group modularity-based fairness)**.** *Let*

$$Q_{\max}^{grp}(S) = \max_{g \in \mathcal{U}} Q_g(S), \qquad Q_{\min}^{grp}(S) = \min_{g \in \mathcal{U}} Q_g(S).$$

We define the normalized multi-group modularity gap, which reports worst–best group-connectivity disparity relative to the overall modularity*, as*

$$\Delta_{\mathcal{U}}(S) = \frac{|Q_{\max}^{grp}(S) - Q_{\min}^{grp}(S)|}{Q(S)}.$$

That is, we normalize the difference between the best and worst connected groups by $Q(S)$ to quantify how large the disparity is compared to the strength of community structure captured by modularity.

To report fairness as a score where higher values indicate more balanced group connectivity, we define the multi-group fairness ratio:

$$F_{\mathcal{U}}(S) = 1 - \Delta_{\mathcal{U}}(S).$$

For $|\mathcal{U}| = 2$, this reduces to the binary modularity-gap formulation.

Multi Deep Group Fairness: We extend DeepFairness by penalizing the gap between the best and worst connected groups. The loss function is:

$$\mathcal{L}_{\text{Multi Deep Group Fairness}} = -Q(S) \; + \; \phi\,(Q_{\max}^{\text{grp}}(S) - Q_{\min}^{\text{grp}}(S)) \; + \; \gamma\, \mathcal{R}_{\text{collapse}},$$

where ϕ controls the strength of the fairness penalty.

4 Opinion Formation Fairness

An opinion-formation model defines a dynamic process by which individuals in a network form opinions. In this part of the project, we propose a novel definition of fairness for opinion formation, that relies on *influence*.

4.1 Opinion Formation Model

The model we consider is the popular Friedkin and Johnsen (FJ) model [8]. In this model, we are given a graph with a set of n nodes V and edges E. Each node $i \in V$ has a fixed *inner opinion* $s_i \in [-1, 1]$ and an *expressed opinion* z_i. The former is fixed, and a characteristic of the node itself; the latter is the result of the opinion formation process that involves the inner opinion of the node and the interaction of the node with the expressed opinions in its social network. Each node i is also associated with a *stubbornness* value $a_i \in (0, 1)$, which determines how opinionated the node is about its inner opinion, and how resistant it is to the opinions of others – higher values of a_i indicate greater stubbornness, meaning that node i places a large weight on their inner opinion and less on the opinion of its social circle.

Each node interacts (iteratively) with its neighboring nodes in the network, adjusting its expressed opinion z_i. These interactions are determined by the *interaction matrix* $W \in [0, 1]^{n \times n}$, which defines a weight w_{ij} for each edge $(i, j) \in E$; w_{ij} determines the importance that node i places on the expressed opinion of neighboring node j. W is row stochastic (i.e., each entry $W[i, j] = w_{ij}$ is non-negative and every row sums to 1).

In FJ, the expressed opinions are updated iteratively. At iteration t the expressed opinion of node i becomes:

$$z_i^{(t)} = a_i s_i + (1 - a_i) \sum_{j \in N_i} w_{ij} z_j^{(t-1)},$$

where N_i is the neighborhood of node i in G. Let $\mathbf{a}$, $\mathbf{s}$ and $\mathbf{z}$ denote the n-dimensional vectors of all stubbornness values, inner and expressed opinions of the nodes in V. We can write the update equation for the FJ model in matrix-vector terms:

$$\mathbf{z}^{(t)} = A\mathbf{s} + (I - A)W\mathbf{z}^{(t-1)}, \tag{11}$$

where $A = \text{Diag}(\mathbf{a})$ is the diagonal matrix with $A[i, i] = a_i$ and I is the $n \times n$ identity matrix. A unique equilibrium vector $\mathbf{z}$ exists if: (i) matrix W is *irreducible* (i.e., the underlying graph is connected) and (ii) at least one node has stubbornness $a_i > 0$. At this steady state, the expressed opinions $\mathbf{z}$ of the nodes in V are:

$$\mathbf{z} = (I - (I - A)W)^{-1} A\mathbf{s} = Q\mathbf{s}. \tag{12}$$

We define the *influence matrix* $Q := (I - (I - A)W)^{-1}A$, which is central to our work. The entries of matrix Q, $Q[i, j] = q_{ij} \in (0, 1)$ determine the *influence* that node j exerts to node i. Note that $z_i = \sum_{j \in V} q_{ij} s_j$, and thus, the value q_{ij} determines the extent to which the expressed opinion z_i of node i is influenced by the inner opinion s_j of node j.

For a node i, we define the *influence of node* i in the network as the average influence node i exerts to all the nodes in the network:

$$Q_i = \frac{1}{n} \sum_{j \in V} q_{ji}. \tag{13}$$

The value Q_i defines the influence of node i to the average opinion $\bar{z} = \frac{1}{n} \sum_{i \in V} z_i$ in the network. This is an important quantity, as it captures the public opinion in the network, and it is often targeted for maximization [9, 1, 33]. We have that $\bar{z} = \sum_{i \in V} Q_i s_i$, thus, Q_i is the influence of the inner opinion s_i to the public opinion.

Note that the matrix Q is row-stochastic, i.e., $\sum_{j \in V} q_{ij} = 1$. Therefore, the total influence exerted by all nodes in the network satisfies: $\sum_{i=1}^{n} Q_i = 1$. This highlights the zero-sum nature of influence in the FJ model: the total amount of influence in the system remains constant and must always sum to one. Consequently, when one node loses influence, the influence of the remaining nodes increases by redistributing the vacated influence among themselves.

4.2 Opinion Fairness

Following the group fairness paradigm, we assume that the nodes in V are partitioned into two groups: *red* ($R \subseteq V$) and *blue* ($B \subseteq V$) with $R \cap B = \emptyset$ and $R \cup B = V$. These groups are defined according to some sensitive attribute such as gender. For the following we use $G = (V, E, W, R, B)$ to denote the weighted graph, where the weights are given by matrix W, and the partition of the nodes into red and blue is given by R and B respectively.

For a group $T \in \{R, B\}$, we define the *group influence* of T as:

$$Q_T = \sum_{i \in T} Q_i, \tag{14}$$

where Q_i is the influence of node i, defined in (13). As discussed, we have that $\sum_{i \in V} Q_i = 1$, and thus $Q_R + Q_B = 1$.

The Q_R and Q_B values determine the strength of the voice of each group within the network and the effect on public opinion. To illustrate, suppose all nodes in each group share the same inner opinion, denoted s_R and s_B for groups R and B. Then, the public opinion is given by $\bar{z} = Q_R s_R + Q_B s_B$. As Q_R increases, $\bar{z}$ shifts toward the Red group's inner opinion s_R, while larger Q_B pushes it toward s_B. For example, if $s_R = 1$ and $s_B = -1$, this simplifies to $\bar{z} = Q_R - Q_B$. In this case, the sign of the network public opinion is determined by the group whose influence exceeds 0.5.

For the opinion formation process to be fair, we require that the Q_R and Q_B values are sufficiently balanced; otherwise, we say that the process is unfair.

Definition 8 (ϕ-Fairness)**.** *Given the input graph $G = (V, E, W, R, B)$, and a parameter $\phi \in (0, 1)$, the FJ opinion-formation process on graph G is ϕ-**fair** if and only if: $Q_R = \phi$.*

The definition of fairness is parameterized by the value $\phi \in (0, 1)$, which enforces different fairness policies. For instance, we can set $\phi = 0.5$ to enforce equal influence between the two groups. Alternatively, we can enforce demographic parity fairness by setting ϕ equal to the fraction of red nodes in the graph. Finally, if the red group is a minority or a protected group, we can also set ϕ to enforce an affirmative action policy, empowering the voice of the minority.

5 k-core fairness

In this part of the project, we consider the novel problem of fairness of the k-core of a graph. The k-core of the graph is defined as follows [24]:

Definition 9 (k-core)**.** *Given a graph G and a positive integer k, an induced subgraph S is the k-core of G if* (i) *$d_S(u) \geq k$ for every vertex $u \in S$ and* (ii) *S is maximal, i.e., any supergraph $S' \supset H$ cannot be a k-core.*

The fairness of k-core has been studied in [34] who give the following fairness definition:

Definition 10 (Fair (k, r)-core)**.** *Given an attributed undirected graph $G = (V, E, A)$, two positive integers k and r, and an induced subgraph $S \subseteq G$. S is a fair (k, r)-core of G, if it satisfies all the following constraints.*

- *Degree: for $\forall v \in V(S), d_S(v) \geq k$.*
- *Fairness: for any two distinct attributes $a_i, a_j \in A$, the difference in the number of vertices with a_i and a_j in S is no larger than the threshold r,*
- *Maximal: any supergraph of S cannot be a fair (k, r)-core.*

In our definitions of k-core fairness, we consider that the attributed graph contains only two attributes, *red* and *blue*. We define the red neighborhood of a vertex v as $NR(v) = \{u \,|\, u \in N(v) \text{ and } u \text{ is red}\}$. We define the blue neighborhood of a vertex v as $NB(v) = \{u \,|\, u \in N(v) \text{ and } u \text{ is blue}\}$. We denote by $d(v)$ the degree of a vertex v and by $dRed(v)$ and $dBlue(v)$ the red and the blue degree of a vertex v respectively, that are $dRed(v) = |NR(v)|$ and $dBlue(v) = |NB(v)|$.

The first definition of k-core fairness we cosnider is the following:

Definition 11 (Fair (k, ℓ)-core)**.** *Given an attributed graph $G = (V, E, A)$, two positive integers k and ℓ, and an induced subgraph $S \subseteq G$. S is a fair (k, ℓ)-core of G, if it satisfies all the following constraints.*

- *Degree: for $\forall v \in V(S), d_S(v) \geq k$.*
- *Fairness: for every vertex v the red and the blue neighborhood in S is at least ℓ, that is $|NR(v)| \geq \ell$ and $|NB(v)| \geq \ell$.*
- *Maximal: any supergraph of S cannot be a fair (k, ℓ)-core.*

We can prove the following:

Proposition 1. *The fair (k, ℓ)-core of a graph is unique.*

Proof. Assume that there are at least two maximal fair (k, ℓ)-cores. Without loss of generality, assume that there are exactly two. We show a contradiction. Let F_1, F_2 both be maximal subgraphs of a graph G such that for every vertex $v \in F_1$ and for every vertex $u \in F_2$ the followings hold:

- $d(v) \geq k$ and $|NR(v)| \geq \ell$ and $|NB(v)| \geq \ell$.
- $d(u) \geq k$ and $|NR(u)| \geq \ell$ and $|NB(u)| \geq \ell$.

Let $F = F_1 \cup F_2$. Then $F_1 \subseteq F$ and $F_1 \subseteq F$. Hence, for every vertex $x \in F$ we have $d(x) \geq k$ and $|NR(x)| \geq \ell$ and $|NB(x)| \geq \ell$. Thus, F is a maximal subgraph that satisfies the degree properties which leads to a contradiction. □

Comparing the (k, r)-core fairness definition with the (k, ℓ)-core fairness definition, we can prove the following:

Proposition 2. *For any fair (k, ℓ)-core, $C_{k,\ell}$, there is a fair (k, r)-core, $C_{k,r}$, such that $C_{k,\ell} \subseteq C_{k,r}$.*

Proof. Let $C_{k,\ell}$ be a fair (k, ℓ)-core with p red and q blue vertices of a graph G. We know that every vertex in $C_{k,\ell}$ has degree at least k. Thus, defining $r \geq |p - q|$ we have that $C_{k,\ell}$ is a fair (k, r)-core (not maximal). A simple example that shows that $C_{k,\ell}$ is not a maximal fair (k, r)-core is the following: Apart from $C_{k,\ell}$ we can consider two vertices, one blue x and one red y, of the graph G such that each of them has at least k neighbors in $C_{k,\ell}$ but without loss of generality $|NR(x)| < \ell$ and $|NB(y)| < \ell$. On the other hand, we can add x and y in $C_{k,\ell}$ and then we construct a larger fair (k, r)-core. □

The opposite it is not true. Let $G = (B \cup R, E)$ be a complete bipartite such that every vertex in B is blue and every vertex in R is red. Then the following holds:

Proposition 3. *Given the colored complete bipartite G, there is a Fair (k, r)-core for some (fixed) values of k and r, although there is no fair (k, ℓ)-core for any value of ℓ.*

The second definition of k-core fairness we consider is the following:

Definition 12 (Fair (k, t)-core)**.** *Given an attributed graph $G = (V, E, A)$, two positive integers k and t, and an induced subgraph $S \subseteq G$. S is a fair (k, t)-core of G, if it satisfies all the following constraints.*

- *Degree: for* $\forall v \in V(S), d_S(v) \geq k$.
- *Fairness: for every vertex* v: $||NR(v)| - |NB(v)|| \leq t$.
- *Maximal: any supergraph of* S *cannot be a fair* (k,t)*-core.*

The fair (k,t)-core is not unique. Consider the graph $G = (V, E)$ with $V = \{a, b, c, d, e\}$ and $E = \{ab, ac, bc, bd, de, de\}$. ($G$ is two triangles with b the common vertex). For $k = 2$ and $t = 1$ there are two maximal fair $(2, 1)$-cores. $C_1 = \{a, b, c\}$ and $C_2 = \{b, d, e\}$.

6 Modeling Homophily in Social Networks

Networks play a central role in many domains, yet the principles that govern their formation are not adequately understood. The goal of this part of the project is to understand the emergence of homophily in Social Networks. Homophily refers to the tendency of social network users to connect with similar users. This is a prevalent property that has been observed in several settings, when nodes are associated with attributes. To model and understand the emergence of homophily, we simulate the creation of a network using *LLM agents*. We vary the attributes of the agents and the mechanism for the link creation and measure the network properties (including homophily). The work was published at the AI for Computational Social Science (AI4CSS) workshop at ICMD 2025 [11].

6.1 Network Creation Algorithm

The network creation process starts with a collection of agents, with no connection between them, that is, an empty network. Agents are assigned attributes that remain fixed throughout the process. At each timestep, an agent selects to connect to another agent from a *candidate pool* of agents. For the selection, we control the information provided to the source agent regarding the attributes of the agents in the pool. For reciprocity, we also model cases where the target agent may reject the connection request. Our overall goal is to understand how attributes shape edge formation and, in turn, the evolution of connectivity, community structure, and homophily in the network.

6.1.1 Agent creation

Agents are created in the timestep $t = 0$ of the simulation process, with no connections between them. Each agent $a \in V$ is assigned attributes specifying its demographics and personality. Each agent is first assigned demographic attributes drawn from predefined distributions. The three attributes considered are sex, race, and age group, as summarized in Table 1, adapted from U.S. Census Bureau American Community Survey (ACS) categories.[3] Sex and race are drawn independently from fixed categorical distributions that approximate population frequencies. The age group is assigned by mapping a normal distribution over adult ages onto the buckets shown in Table 1.

In addition to demographic attributes, each agent is assigned *Big Five trait scores* (OCEAN) [14, 32]: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. These traits provide a continuous representation of individual differences in behavior and social preferences. In our simulation, values for each trait are sampled from a truncated normal distribution in $[0, 1]$, independent of demographics, to ensure variation while remaining within plausible ranges.

Demographic and psychological attributes together serve as inputs for connection decisions. Demographics ensure that population diversity is explicitly represented in the simulation, while the Big Five traits introduce heterogeneous decision-making patterns, influencing both the number of connections agents attempt and the likelihood of accepting or rejecting connection requests.

[3] See https://www.census.gov.

Table 1: Demographic attribute values used in simulation

Attribute	Values
Sex	Male, Female
Race	White, Black, Asian, Hispanic
Age Group	18–24, 25–34, 35–44, 45–54, 55–64, 65+

Finally, each agent a maintains a persistent rejection memory $M_t(a) \subseteq V \setminus \{a\}$ that stores the identities of agents that declined its connection requests in the past.

6.1.2 Network creation

Let $G_t = (V, E_t)$ denote the cumulative undirected graph after timestep t, where V is the set of agents and E_t the set of edges where each edge is a timestamped tuple (i, j, τ), $i, j \in V$, and $\tau \leq t$. At $t = 0$, we create n agents and assign to them fixed attributes that persist for the entire simulation. There are no edges between them, so E_0 is empty.

The process of creating connections proceeds as follows. At timestep t, each agent a initiates k connection attempts from a candidate pool of agents. The candidate pool for agent a consists of all second-hop (friend-of-friend) nodes, plus k nodes sampled uniformly at random, excluding the self node a, current neighbors $N_t(a)$, and prior rejections $M_t(a)$.

Agent a maintains a persistent rejection memory $M_t(a) \subseteq V$ with the identities of agents that rejected a in the past. If a proposes to b at step t and b rejects, we update $M_t(a) \leftarrow M_t(a) \cup \{b\}$. The memory records identities only and does not include time, context, or reasons. At every step the candidate pool excludes prior rejectors.

The model selects up to k targets from the candidate pool. We use the same k for sampling and selection.

6.1.3 Prompt design

We now provide details about the two templates we use: a *selection* prompt and an *acceptance* prompt. Both prompts are parameterized by the current timestep t, the focal agent a, and the candidate pool $C_t(a)$.

- **Selection prompt.** The prompt includes: (1) a header with agent a's own demographics and Big Five traits (2) a list of candidates $u \in C_t(a)$ that includes an anonymized ID (for example, `U17`), whether u is a friend of friend (FoF) or a random draw and any attributes, and (3) the instruction that asks for up to k different targets from $C_t(a)$ in a fixed output format.
- **Acceptance prompt.** The prompt includes: (1) a header with target b's own demographics and Big Five traits, (2) a initiator summary that shows a's anonymized ID (e.g., `U17`), whether a is a FoF or a random draw relative to b and any attributes, and (3) the instruction that requests a binary decision (ACCEPT or REJECT) with a one-sentence rationale. If the decision is REJECT, b is added to $M_t(a)$ so that a does not propose to b again.

All runs use a fixed prompt template for reproducibility.

6.2 Evaluation

We examine how demographic attributes and Big Five traits shape the selection of connections. We evaluate an agent driven process in which each agent selects a target from a candidate pool, and, when acceptance is on, the selected target decides whether to accept the proposal. To assess the

role of agent characteristics in link selection, we vary what is visible to the agents by revealing or hiding demographics (age, sex, race) and Big Five traits. To evaluate the impact of reciprocity, we toggle acceptance so that a proposed edge is added only if the target agrees. We run the selection process in multiple timesteps to see how connections evolve.

A central finding concerns the strong emergence of homophily when demographic attributes are visible. When agents can observe demographics, same-attribute connections increase dramatically by the final snapshot: race homophily nearly doubles and age homophily triples compared to settings where demographics are hidden. These gains are accompanied by substantially higher modularity and more fragmented community structure, indicating that demographic visibility leads to tightly knit, demographically aligned communities with fewer cross-group links. In contrast, when demographics are hidden, homophily remains close to baseline levels, and the resulting networks are more mixed and less modular.

Visibility of Big Five personality traits also induces homophily, though in a more nuanced way. When traits are visible and demographics are hidden, agents preferentially connect to others with similar personality profiles, with the strongest effects observed for Openness and Extraversion. Trait-driven homophily increases across all five traits, but the resulting community structure is less segregated than in the demographic case: modularity slightly decreases, suggesting that personality similarity creates cross-cutting ties that span communities rather than reinforcing strict partitions. When both demographics and traits are visible, trait homophily remains strong, but demographic homophily—especially for race and age—is reduced, indicating that personality information partially counteracts demographic clustering.

Finally, reciprocity through acceptance primarily acts as a secondary filter rather than the main driver of homophily. Enabling acceptance reduces the total number of edges and increases modularity and separation, but it leads only to modest increases in both demographic and trait homophily. This suggests that candidate selection, rather than acceptance, is the dominant mechanism shaping homophilous structure. Compared to Erdős–Rényi and Barabási–Albert baselines matched for size and density, the agent-driven networks consistently exhibit higher modularity, confirming that agent attributes and decision-making play a key role in the emergence of homophily and community structure.

Our results have dual significance. They clarify the mechanisms that drive social network formation and they reveal systematic preferences of current LLMs that can disadvantage groups. Because edges arise from model decisions, these preferences become structural and translate into visibility gaps, degree differences, and community segregation. We treat this framework as a testbed for fairness, and will design and evaluate interventions that counter bias, including adjustments to candidate pool composition, masking or balancing attribute visibility, calibrating acceptance utilities, and adding exposure or degree constraints. Our goal is to design mechanisms that improve fairness while preserving connectivity and community cohesion.

7 Conclusions

he objectives of Work Package 1 were threefold: (1) to investigate metrics used for measuring bias and fairness across different contexts; (2) to propose new metrics for fairness and bias in underexplored problem settings; and (3) to develop models for interpreting the emergence of bias. In this report, we focused on the latter two objectives, presenting novel metrics for bias and fairness as well as models for understanding how bias arises.

With respect to metrics, our contributions are the following:

- We introduce a novel metric that quantifies the cost of unfairness in counterfactual generation.
- We propose new metrics for community fairness and diversity grounded in modularity.
- We define a novel influence-based metric for fairness in opinion formation processes.
- We present alternative formulations of fair k-cores based on edge-level fairness.

Regarding models, we investigate the emergence of homophily in social networks through an LLM-based agent simulation framework. Our analysis demonstrates how the visibility of node attributes influences link formation and contributes to homophily in networks.

Overall, this work lays the foundation for the design of fair algorithms and the formulation of new research questions in responsible AI. Algorithms that optimize or leverage the proposed metrics are developed and discussed in Deliverable 3.4.

Acknowledgements

We would like to acknowledge the contributions of Nikos Theologis and Evimaria Terzi in this work.

References

[1] Rediet Abebe et al. "Opinion Dynamics Optimization by Varying Susceptibility to Persuasion via Non-Convex Local Search". In: *ACM Trans. Knowl. Discov. Data* (2021).

[2] Emilio Carrizosa, Jasone Ramírez-Ayerbe, and Dolores Romero Morales. "Mathematical optimization modelling for group counterfactual explanations". In: *European Journal of Operational Research* (2024).

[3] Matteo Ceccarello, Andrea Pietracaprina, and Geppino Pucci. "Fast and Accurate Fair k-Center Clustering in Doubling Metrics". In: *WWW*. 2024.

[4] Flavio Chierichetti et al. "Fair clustering through fairlets". In: *Advances in neural information processing systems* (2017).

[5] A. Clauset, M. E. J. Newman, and C. Moore. "Finding community structure in very large networks". In: *Phys. Rev. E* 70 (6 2004).

[6] David A. Easley and Jon M. Kleinberg. *Networks, Crowds, and Markets - Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[7] Santo Fortunato. "Community detection in graphs". In: *Physics Reports* 486.3–5 (Feb. 2010), pp. 75–174. ISSN: 0370-1573.

[8] Noah E. Friedkin and Eugene C. Johnsen. "Social influence and opinions". In: *The Journal of Mathematical Sociology* 15.3-4 (1990), pp. 193–206. DOI: 10.1080/0022250X.1990.9990069.

[9] Aristides Gionis, Evimaria Terzi, and Panayiotis Tsaparas. "Opinion Maximization in Social Networks". In: *Proceedings of the 2013 SIAM International Conference on Data Mining (SDM)*. Philadelphia, PA, USA: SIAM, 2013, pp. 387–395. DOI: 10.1137/1.9781611972832.43. URL: https://epubs.siam.org/doi/abs/10.1137/1.9781611972832.43.

[10] Christos Gkartzios, Evaggelia Pitoura, and Panayiotis Tsaparas. "Fair Network Communities through Group Modularity". In: *Proceedings of the ACM Web Conference (WWW '25)*. 2025.

[11] Christos Gkartzios, Evaggelia Pitoura, and Panayiotis Tsaparas. "Modeling Network Formation with LLM Agents: The Role of Demographics and Personality". In: *Workshop on AI for Computational Social Science (AI4CSS), IEEE International Conference on Data Mining*. 2025.

[12] Christos Gkartzios, Evaggelia Pitoura, and Panayiotis Tsaparas. "Modularity-Fair Deep Community Detection". In: *ICDM*. 2025.

[13] Riccardo Guidotti. "Counterfactual explanations and how to find them: literature review and benchmarking". In: *Data Mining and Knowledge Discovery* (2022), pp. 1–55.

[14] Oliver P. John, Laura P. Naumann, and Christopher J. Soto. "Paradigm Shift to the Integrative Big-Five Trait Taxonomy". In: *Handbook of Personality: Theory and Research*. Ed. by O. P. John, R. W. Robins, and L. A. Pervin. Guilford Press, 2008, pp. 114–158.

[15] Amir-Hossein Karimi et al. “A survey of algorithmic recourse: contrastive explanations and consequential recommendations”. In: *ACM Computing Surveys* 55.5 (2022), pp. 1–29.

[16] Amir-Hossein Karimi et al. “Model-agnostic counterfactual explanations for consequential decisions”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 895–905.

[17] Loukas Kavouras et al. “Fairness Aware Counterfactuals for Subgroups”. In: *Advances in Neural Information Processing Systems* 36 (2024).

[18] Matthäus Kleindessner et al. “Guarantees for spectral clustering with fairness constraints”. In: *International conference on machine learning*. 2019, pp. 3458–3467.

[19] Alejandro Kuratomi et al. “CounterFair: Group Counterfactuals for Bias Detection, Mitigation and Subgroup Identification”. In: *IEEE International Conference on Data Mining (ICDM)*. 2024.

[20] Alejandro Kuratomi et al. “Measuring the Burden of (Un) fairness Using Counterfactuals”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2022, pp. 402–417.

[21] Alejandro Kuratomi et al. “Subgroup fairness based on shared counterfactuals”. In: *Knowl. Inf. Syst.* 67.11 (2025), pp. 10863–10901. DOI: 10.1007/S10115-025-02555-7.

[22] Matt J Kusner et al. “Counterfactual fairness”. In: *Advances in neural information processing systems* 30 (2017).

[23] Jure Leskovec, Anand Rajaraman, and Jeffrey D. Ullman. *Mining of Massive Datasets, 2nd Ed.* Cambridge University Press, 2014.

[24] Fragkiskos D Malliaros et al. “The core decomposition of networks: Theory, algorithms and applications”. In: *The VLDB Journal* 29.1 (2020), pp. 61–92.

[25] Ninareh Mehrabi et al. “A Survey on Bias and Fairness in Machine Learning”. In: *ACM Comput. Surv.* 54.6 (2022), 115:1–115:35.

[26] Christoph Molnar. *Interpretable Machine Learning: A Guide for Making black-box Models Explainable*. 2021. URL: https://christophm.github.io/interpretable-ml-book/limo.html.

[27] M. E. J. Newman. “Fast algorithm for detecting community structure in networks”. In: *Phys. Rev. E* 69 (2004).

[28] M. E. J. Newman. “Finding community structure in networks using the eigenvectors of matrices”. In: *Phys. Rev. E* 74 (3 Sept. 2006), p. 036104. DOI: 10.1103/PhysRevE.74.036104.

[29] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. “Fairness in rankings and recommendations: an overview”. In: *VLDB J.* 31.3 (2022), pp. 431–458.

[30] Akrati Saxena, George Fletcher, and Mykola Pechenizkiy. “Fairsna: Algorithmic fairness in social network analysis”. In: *CSUR* 56.8 (2024), pp. 1–45.

[31] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. “CERTIFAI: Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence models”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (Feb. 2020). arXiv: 1905.07857, pp. 166–172. DOI: 10.1145/3375627.3375812. URL: http://arxiv.org/abs/1905.07857 (visited on 03/05/2022).

[32] Christopher J. Soto and Oliver P. John. “The Next Big Five Inventory (BFI-2): Developing and Assessing a Hierarchical Model with 15 Facets to Enhance Bandwidth, Fidelity, and Predictive Power”. In: *Journal of Personality and Social Psychology* 113.1 (2017), pp. 117–143.

[33] Haoxin Sun and Zhongzhi Zhang. “Opinion optimization in directed social networks”. In: *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI’23/IAAI’23/EAAI’23. AAAI Press, 2023. ISBN: 978-1-57735-880-0.

[34] Xingyu Tan et al. “Maximum Fairness-Aware (k, r)-Core Identification in Large Graphs”. In: *Databases Theory and Applications*. Ed. by Zhifeng Bao et al. Cham: Springer Nature Switzerland, 2024, pp. 273–286. ISBN: 978-3-031-47843-7.

[35] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. *Social Media Mining: An Introduction*. Cambridge University Press, 2014.