Networked Inequality: Preferential Attachment Bias in Graph Neural Network Link Prediction

Arjun Subramonian [1] **Levent Sagun** [2] **Yizhou Sun** [1]

Abstract

Graph neural network (GNN) link prediction is increasingly deployed in citation, collaboration, and online social networks to recommend academic literature, collaborators, and friends. While prior research has investigated the dyadic fairness of GNN link prediction, the within-group (e.g., queer women) fairness and "rich get richer" dynamics of link prediction remain underexplored. However, these aspects have significant consequences for degree and power imbalances in networks. In this paper, we shed light on how degree bias in networks affects Graph Convolutional Network (GCN) link prediction. In particular, we theoretically uncover that GCNs with a symmetric normalized graph filter have a within-group preferential attachment bias. We validate our theoretical analysis on real-world citation, collaboration, and online social networks. We further bridge GCN's preferential attachment bias with unfairness in link prediction and propose a new within-group fairness metric. This metric quantifies disparities in link prediction scores within social groups, towards combating the amplification of degree and power disparities. Finally, we propose a simple training-time strategy to alleviate within-group unfairness, and we show that it is effective on citation, social, and credit networks.

1. Introduction

Link prediction (LP) using GNNs is increasingly leveraged to recommend friends in social networks (Fan et al., 2019; Sankar et al., 2021), as well as by scholarly tools to recommend academic literature in citation networks (Xie et al., 2021). In recent years, graph learning researchers have raised concerns about the unfairness of GNN LP (Li et al., 2021; Current et al., 2022; Li et al., 2022). This unfairness is often attributed to graph structure, including the stratification of social groups; for example, online networks are usually segregated by ethnicity (Hofstra et al., 2017). However, most fair GNN LP research has focused on dyadic fairness, i.e., satisfying some notion of parity between inter-group and intra-group link predictions. This formulation neglects: 1) LP dynamics within social groups (Kasy & Abebe, 2021); and 2) the "rich get richer" effect, i.e., the prediction of links at a higher rate with high-degree nodes (Barabási & Albert, 1999). In the context of friend recommendation systems, the "rich get richer" effect can increase the number of links formed with high-degree individuals, which boosts their influence on other individuals in the network, and thus their power (Bashardoust et al., 2022).

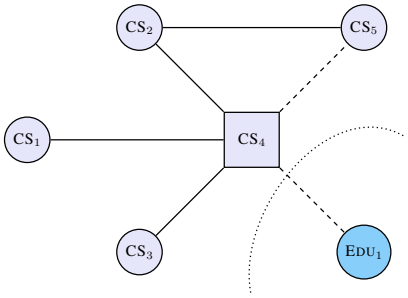


Figure 1: An academic collaboration network where nodes are Computer Science (CS) and Education (EDU) researchers, solid edges are current or past collaborations, and dashed edges are collaborations recommended by a GCN. Circular nodes are women and square nodes are men.

In this paper, we shed light on how degree bias in networks affects GCN LP (Kipf & Welling, 2017). We theoretically and empirically find that GCNs with a symmetric normalized graph filter have a within-group preferential attachment (PA) bias in LP. Specifically, GCNs often output LP scores that are approximately proportional to the geometric mean of the (within-group) degrees of the incident nodes when the nodes belong to the same social group. (We elaborate on PA and our motivation in §J.) We focus on GCNs with symmetric and random walk normalized graph filters because they

*Equal contribution [1]Computer Science Department, University of California, Los Angeles, USA [2]Meta, Paris, France. Correspondence to: Arjun Subramonian <arjunsub@cs.ucla.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

are popular architectures for graph deep learning, and they provide us with a reasonable setting to develop a rigorous theory of PA bias in GNN LP while leveraging tools from spectral graph theory.

Our finding can have significant implications for the fairness of GCN LP. For example, consider links within the CS social group in the toy academic collaboration network in Figure 1. Because men in CS, on average, have a higher within-group degree (deg $= 3$) than women in CS (deg $= 1.25$), due to gender discrimination, a collaboration recommender system that uses a GCN can suggest men as collaborators at a higher rate. This has the detrimental effect of further concentrating research collaborations among men, thereby reducing the influence of women in CS and reinforcing their marginalization in the field (Yamamoto & Frachtenberg, 2022). Furthermore, considering this marginalization in the context of CS is important, as such marginalization may be less severe or different in EDU.

Our contributions are as follows:

1. We theoretically uncover that GCNs with a symmetric normalized graph filter have a within-group PA bias in LP (§4.1). We validate our theoretical analysis on diverse real-world network datasets (e.g., citation, collaboration, online social networks) of varying size (§6.1). In doing so, we lay a foundation to study this previously-unexplored PA bias in the GNN setting.
2. We bridge GCN's PA bias with unfairness in LP (§4.2, §6.2). We contribute a new within-group fairness metric for LP, which quantifies disparities in LP scores within social groups, towards combating the amplification of degree and power disparities. To our knowledge, we are the first to study the within-group fairness of GNNs.
3. We propose a training-time strategy to alleviate within-group unfairness (§5), and we assess its effectiveness on citation, online social, and credit networks (§6.3). Our experiments reveal that even for this new form of unfairness, simple regularization approaches can be successful.

2. Related Work

Degree Bias in GNNs Numerous papers have investigated how GNN performance is degraded for low-degree nodes on node representation learning and classification tasks (Tang et al., 2020; Liu et al., 2021; Kang et al., 2022; Xu et al., 2023; Shomer et al., 2023). Liu et al. (2023) present a generalized notion of degree bias that considers different multi-hop structures around nodes and propose a framework to address it; in contrast to prior work, which focuses on *degree equal opportunity* (i.e., similar accuracy for nodes with the same degree), Liu et al. (2023) also study *degree statistical parity* (i.e., similar prediction rates of each class for nodes with the same degree). Beyond node classification, Wang & Derr (2022) find GNN LP performance disparities across nodes with different degrees: low-degree nodes often benefit from higher performance than high-degree nodes. In this paper, we find that GCNs have a PA bias in LP, and present a new fairness metric which quantifies disparities in GNN LP scores within social groups. We focus on *group fairness* (i.e., parity between groups) rather than *individual fairness* (i.e., treating similar individuals similarly); this is because producing similar LP scores for similar-degree individuals does not prevent high-degree individuals from unfairly amassing links, and thus power (cf. Figure 1). We further compare our work to prior degree bias works in §K.

Fair Link Prediction Prior work has investigated the unfairness of GNN LP (Li et al., 2021; Current et al., 2022; Li et al., 2022), often attributing it to graph structure, (e.g., stratification of social groups). However, most of this research has focused on dyadic fairness, i.e., satisfying some notion of parity between inter-group and intra-group links. Like Wang & Derr (2022), we examine how degree bias impacts GNN LP; however, rather than focus on performance disparities across nodes with different degrees, we study GCN's PA bias and LP score disparities across (sub)groups.

Within-Group Fairness Much previous work has studied within-group fairness, i.e., fairness over social subgroups (e.g., Black women, Indigenous men) defined over multiple axes (e.g., race, gender) (Kearns et al., 2017; Foulds et al., 2020; Ghosh et al., 2021; Wang et al., 2022). The motivation of this work is that classifiers can be fair with respect to two social axes separately, but be unfair to subgroups defined over both these axes. While prior research has termed this phenomenon *intersectional* unfairness, we opt for *within-group* unfairness to distinguish it from the critical framework of Intersectionality (Ovalle et al., 2023). We study within-group fairness in the GNN setting. In particular, our theoretical and empirical findings reveal that GCN LP can further marginalize social subgroups; this relates to the "complexity" tenet of Intersectionality, which expresses that the marginalization faced by, e.g., Black women, is non-additive and distinct from the marginalization faced by Black men and white women (Collins & Bilge, 2020).

Bias and Power in Networks A wealth of literature outside fair graph learning has examined how network structure enables discrimination and disparities in capital (Fish et al., 2019; Stoica et al., 2020; Zhang et al., 2021; Bashardoust et al., 2022). Boyd et al. (2014) describe how an individual's position in a social network affects their access to jobs and public health information, as well as how they are surveilled. Stoica et al. (2018) observe that high-degree accounts on Instagram overwhelmingly belong to men and recommendation algorithms further boost these accounts; complementarily, the authors find that even a simple, ran-

dom walk-based recommendation algorithm can amplify degree disparities between social groups in networks modeled by PA dynamics. Similarly, we investigate how GCN LP can amplify degree disparities in networks and further concentrate power among high-degree individuals.

3. Preliminaries

We have a simple, undirected n-node graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with doubly-weighted self-loops. The nodes have features $(\boldsymbol{x}_i)_{i\in\mathcal{V}}$, with each $\boldsymbol{x}_i \in \mathbb{R}^d$. We denote the adjacency matrix of $\mathcal{G}$ as $\boldsymbol{A} \in \{0,1\}^{n\times n}$ and the degree matrix as $\boldsymbol{D} = \text{diag}\left(\left(\sum_{j\in\mathcal{V}} \boldsymbol{A}_{ij}\right)_{i\in\mathcal{V}}\right)$, with $\boldsymbol{D} \in \mathbb{N}^{n\times n}$. We consider two L-layer GCN encoders: (1) $\Phi_s : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d'}$ (Kipf & Welling, 2017), which uses a symmetric normalized filter, and (2) $\Phi_r : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times d'}$, which uses a random walk normalized filter. Φ_s and Φ_r compute node representations as, $\forall i \in \mathcal{V}$:

$$\Phi_s\left((\boldsymbol{x}_j)_{j\in\mathcal{V}}\right)_i = \boldsymbol{s}_i^{(L)}, \Phi_r\left((\boldsymbol{x}_j)_{j\in\mathcal{V}}\right)_i = \boldsymbol{r}_i^{(L)} \quad (1)$$

$$\forall l \in [L], \boldsymbol{s}_i^{(l)} = \sigma^{(l)}\left(\sum_{j\in\Gamma(i)} \frac{\boldsymbol{W}_s^{(l)}\boldsymbol{s}_j^{(l-1)}}{\sqrt{\boldsymbol{D}_{ii}\boldsymbol{D}_{jj}}}\right), \quad (2)$$

$$\forall l \in [L], \boldsymbol{r}_i^{(l)} = \sigma^{(l)}\left(\sum_{j\in\Gamma(i)} \frac{\boldsymbol{W}_r^{(l)}\boldsymbol{r}_j^{(l-1)}}{\boldsymbol{D}_{ii}}\right), \quad (3)$$

where $\left(\boldsymbol{s}_i^{(0)}\right)_{i\in\mathcal{V}} = \left(\boldsymbol{r}_i^{(0)}\right)_{i\in\mathcal{V}} = (\boldsymbol{x}_i)_{i\in\mathcal{V}}$; $\Gamma(i)$ is the 1-hop neighborhood of i; $\boldsymbol{W}_s^{(l)}$ and $\boldsymbol{W}_r^{(l)}$ are the weight matrices corresponding to layer l of Φ_s and Φ_r, respectively; for $l \in [L-1]$, $\sigma^{(l)}$ is a ReLU non-linearity; and $\sigma^{(L)}$ is the identity function. We now consider the first-order Taylor expansions of Φ_s and Φ_r around $(\boldsymbol{0})_{i\in\mathcal{V}}$:

$$\boldsymbol{s}_i^{(L)} = \sum_{j\in\mathcal{V}} \left[\frac{\partial \boldsymbol{s}_i^{(L)}}{\partial \boldsymbol{x}_j}\right] \boldsymbol{x}_j + \xi\left(\boldsymbol{s}_i^{(L)}\right), \quad (4)$$

$$\boldsymbol{r}_i^{(L)} = \sum_{j\in\mathcal{V}} \left[\frac{\partial \boldsymbol{r}_i^{(L)}}{\partial \boldsymbol{x}_j}\right] \boldsymbol{x}_j + \xi\left(\boldsymbol{r}_i^{(L)}\right), \quad (5)$$

where ξ is the error of the first-order approximations. This error is low when $(\boldsymbol{x}_i)_{i\in\mathcal{V}}$ are close to $\boldsymbol{0}$, which we validate empirically in §6.1. Furthermore, we consider an inner-product LP score function $f_{LP} : \mathbb{R}^{d'} \times \mathbb{R}^{d'} \to \mathbb{R}$:

$$f_{LP}\left(\boldsymbol{h}_i^{(L)}, \boldsymbol{h}_j^{(L)}\right) = \left(\boldsymbol{h}_i^{(L)}\right)^\intercal \boldsymbol{h}_j^{(L)}, \quad (6)$$

where $\boldsymbol{h}_i^{(L)}$ is the last-layer representation for node i. While it is common to use a vanilla GCN and inner-product score function for LP (Fey, 2019), researchers have proposed methods to improve the expressivity of node representations for LP by capturing subgraph information (Zhang & Chen, 2018; Li et al., 2020; Chamberlain et al., 2023). Our theoretical findings remain relevant to methods that ultimately use a GCN to predict links (e.g., Zhang & Chen (2018); Li et al. (2020)), as we do not make assumptions about the features passed to the GCN (i.e., they could be distance encodings, SEAL node embeddings, etc.) Our results may also generalize to GNN architectures that use a degree-normalized graph filter, e.g., Graph Attention Networks (Veličković et al., 2018). Studying the fairness of more expressive LP methods is an interesting direction for future research. Furthermore, although we only consider an inner-product LP score function in our theoretical analysis, we also run experiments with a Hadamard product and MLP score function (cf. §G.2), and we find that our theoretical analysis is still relevant to and reasonably supports the experimental results.

4. Theoretical Analysis

We leverage spectral graph theory to study how degree bias affects GCN LP. Theoretically, we find that GCNs with a symmetric normalized graph filter have a within-group PA bias (§4.1), but GCNs with a random walk normalized filter may lack such a bias (§4.3). We further bridge GCN's PA bias with unfairness in GCN LP, proposing a new LP within-group fairness metric (§4.2) and a simple training-time strategy to alleviate unfairness (§5). We empirically validate our theoretical results and fairness strategy in §6. We provide proofs for all theoretical results in §A.

Our ultimate goal is to bound the expected LP scores $\mathbb{E}\left[f_{LP}\left(\boldsymbol{s}_i^{(L)}, \boldsymbol{s}_j^{(L)}\right)\right]$ and $\mathbb{E}\left[f_{LP}\left(\boldsymbol{r}_i^{(L)}, \boldsymbol{r}_j^{(L)}\right)\right]$ for nodes i, j in the same social group in terms of the degrees of i, j. We begin with Lemma 4.1, which expresses GCN representations (in expectation) as a linear combination of the initial node features. In doing so, we decouple the computation of GCN representations from the non-linearities $\sigma^{(l)}$.

Lemma 4.1. *Similarly to Xu et al. (2018), assume that each path from node $i \to j$ in the computation graph of Φ_s is independently activated with probability $\rho_s(i)$, and similarly, $\rho_r(i)$ for Φ_r (cf. §L). Furthermore, suppose that $\mathbb{E}\left[\xi\left(\boldsymbol{s}_i^{(L)}\right)\right] = \mathbb{E}\left[\xi\left(\boldsymbol{r}_i^{(L)}\right)\right] = \boldsymbol{0}$, where the expectations are taken over the probability distributions of paths activating. We define $\alpha_j = \left(\prod_{l=L}^1 \boldsymbol{W}_s^{(l)}\right)\boldsymbol{x}_j$, and $\beta_j = \left(\prod_{l=L}^1 \boldsymbol{W}_r^{(l)}\right)\boldsymbol{x}_j$. Then, $\forall i \in \mathcal{V}$:*

$$\mathbb{E}\left[\boldsymbol{s}_i^{(L)}\right] = \sum_{j\in\mathcal{V}} \rho_s(i)\left(\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}\right)_{ij}^L \alpha_j, \quad (7)$$

$$\mathbb{E}\left[\boldsymbol{r}_i^{(L)}\right] = \sum_{j\in\mathcal{V}} \rho_r(i)\left(\boldsymbol{D}^{-1}\boldsymbol{A}\right)_{ij}^L \beta_j. \quad (8)$$

Lemma 4.1 demonstrates that under certain assumptions

(which we show to be reasonable in §6.1), the expected GCN representations can be expressed as a linear combination of the node features that depends on a normalized version of the adjacency matrix.

We now introduce social groups in $\mathcal{G}$ into our analysis. Suppose that $\mathcal{V}$ can be partitioned into B disjoint social groups $\{S^{(b)}\}_{b\in[B]}$, such that $\bigcup_{b\in[B]} S^{(b)} = \mathcal{V}$ and $\bigcap_{b\in[B]} S^{(b)} = \emptyset$. Furthermore, we define $\mathcal{G}^{(b)}$ as the induced connected subgraph of $\mathcal{G}$ formed from $S^{(b)}$. (If a group comprises $C > 1$ connected components, it can be treated as C separate groups.) Let $\widehat{\boldsymbol{A}}$ be a within-group adjacency matrix that contains links between nodes in the same group, i.e., $\widehat{\boldsymbol{A}}$ contains the link (i, j) if and only if for some group $S^{(b)}$, $i, j \in S^{(b)}$. Without loss of generality, we reorder the rows and columns of $\widehat{\boldsymbol{A}}$ and $\boldsymbol{A}$ such that $\widehat{\boldsymbol{A}}$ is a block matrix. Let $\widehat{\boldsymbol{D}}$ be the degree matrix of $\widehat{\boldsymbol{A}}$.

4.1. Symmetric Normalized Filter

We first focus on analyzing Φ_s. We introduce the notation $\boldsymbol{P} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}$ for the symmetric normalized adjacency matrix. We further define $\widehat{\boldsymbol{P}} = \widehat{\boldsymbol{D}}^{-\frac{1}{2}}\widehat{\boldsymbol{A}}\widehat{\boldsymbol{D}}^{-\frac{1}{2}}$, which has the form $\begin{bmatrix} \widehat{\boldsymbol{P}}^{(1)} & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \widehat{\boldsymbol{P}}^{(B)} \end{bmatrix}$. Each $\widehat{\boldsymbol{P}}^{(b)}$ admits the orthonormal spectral decomposition $\widehat{\boldsymbol{P}}^{(b)} = \sum_{k=1}^{\left|S^{(b)}\right|} \lambda_k^{(b)} \boldsymbol{v}_k^{(b)} \left(\boldsymbol{v}_k^{(b)}\right)^{\intercal}$. Let $\left(\lambda_k^{(b)}\right)_{1\le k\le \left|S^{(b)}\right|}$ be the eigenvalues of $\widehat{\boldsymbol{P}}^{(b)}$ sorted in non-increasing order; the eigenvalues fall in the range $(-1, 1]$. By the spectral properties of $\widehat{\boldsymbol{P}}^{(b)}$, $\lambda_1^{(b)} = 1$. Following Lovász (2001), we denote the *spectral gap* of $\widehat{\boldsymbol{P}}^{(b)}$ as $\lambda^{(b)} = \max\left\{\lambda_2^{(b)}, \left|\lambda_{\left|S^{(b)}\right|}^{(b)}\right|\right\} < 1$; $\lambda_2^{(b)}$ corresponds to the smallest non-zero eigenvalue of the symmetric normalized graph Laplacian. Let $\boldsymbol{P} = \widehat{\boldsymbol{P}} + \Xi^{(0)}$. If $\mathcal{G}$ is highly modular or approximately disconnected, then $\Xi^{(0)} \approxeq \boldsymbol{0}$, albeit with positive and non-positive entries. Finally, we define the volume $\text{vol}\left(\mathcal{G}^{(b)}\right) = \sum_{k\in S^{(b)}} \widehat{\boldsymbol{D}}_{kk}$.

In Lemma 4.2, we present an inequality for the entries of $\boldsymbol{P}^L$ in terms of the spectral properties of $\widehat{\boldsymbol{P}}$. We can then combine this inequality with Lemma 4.1 to bound $\mathbb{E}\left[\boldsymbol{s}_i^{(L)}\right]$, and subsequently $\mathbb{E}\left[f_{LP}\left(\boldsymbol{s}_i^{(L)}, \boldsymbol{s}_j^{(L)}\right)\right]$.

Lemma 4.2. *For $i, j \in S^{(b)}$:*

$$\left|\boldsymbol{P}_{ij}^L - \frac{\sqrt{\widehat{\boldsymbol{D}}_{ii}\widehat{\boldsymbol{D}}_{jj}}}{vol\left(\mathcal{G}^{(b)}\right)}\right| \tag{9}$$

$$\le \zeta_s = \left(\lambda^{(b)}\right)^L + \sum_{l=1}^{L}\binom{L}{l}\left\|\Xi^{(0)}\right\|_{op}^{l}\left\|\widehat{\boldsymbol{P}}\right\|_{op}^{L-l}, \tag{10}$$

where $\|\cdot\|_{op}$ is the operator norm. And for $i \in S^{(b)}, j \notin S^{(b)}$, $\left|\boldsymbol{P}_{ij}^L - 0\right| \le \sum_{l=1}^{L}\binom{L}{l}\left\|\Xi^{(0)}\right\|_{op}^{l}\left\|\widehat{\boldsymbol{P}}\right\|_{op}^{L-l} \le \zeta_s$.

The proof of Lemma 4.2 is similar to spectral proofs of random walk convergence. When L is small (e.g., 2 for many GCNs (Kipf & Welling, 2017)) and $\left\|\Xi^{(0)}\right\|_{op} \approxeq 0$, $\sum_{l=1}^{L}\binom{L}{l}\left\|\Xi^{(0)}\right\|_{op}^{l}\left\|\widehat{\boldsymbol{P}}\right\|_{op}^{L-l} \approxeq 0$. Furthermore, with significant stratification between social groups (Hofstra et al., 2017) and high expansion within groups (Malliaros & Megalooikonomou, 2011; Leskovec et al., 2008), $\lambda^{(b)} << 1$. In this case, $\zeta_s \approxeq 0$ and $\boldsymbol{P}_{ij}^L \approxeq \frac{\sqrt{\widehat{\boldsymbol{D}}_{ii}\widehat{\boldsymbol{D}}_{jj}}}{\text{vol}\left(\mathcal{G}^{(b)}\right)}$ for $i, j \in S^{(b)}$. Combining Lemmas 4.1 and 4.2, Φ_s can oversmooth the expected representations to $\mathbb{E}\left[\boldsymbol{s}_i^{(L)}\right] \approxeq \rho_s(i)\sqrt{\widehat{\boldsymbol{D}}_{ii}} \cdot \sum_{j\in S^{(b)}} \frac{\sqrt{\widehat{\boldsymbol{D}}_{jj}}}{\text{vol}\left(\mathcal{G}^{(b)}\right)}\alpha_j$ (Keriven, 2022; Giovanni et al., 2023). We use this knowledge to bound $\mathbb{E}\left[f_{LP}\left(\boldsymbol{s}_i^{(L)}, \boldsymbol{s}_j^{(L)}\right)\right]$ in terms of the degrees of i, j.

Theorem 4.3. *Following a relaxed assumption from Xu et al. (2018), for nodes $i, j \in S^{(b)}$, we assume that $\rho_s(i) = \rho_s(j) = \overline{\rho}_s(b)$. Then:*

$$\left|\mathbb{E}\left[f_{LP}\left(\boldsymbol{s}_i^{(L)}, \boldsymbol{s}_j^{(L)}\right)\right] - C_0\sqrt{\widehat{\boldsymbol{D}}_{ii}\widehat{\boldsymbol{D}}_{jj}}\right| \tag{11}$$

$$\le \zeta_s\overline{\rho}_s^2(b)\left(\sqrt{\widehat{\boldsymbol{D}}_{ii}} + \sqrt{\widehat{\boldsymbol{D}}_{jj}}\right)C_1C_2 + \zeta_s^2\overline{\rho}_s^2(b)C_2^2, \tag{12}$$

where: (13)

$$C_0 = \overline{\rho}_s^2(b)C_1^2, \tag{14}$$

$$C_1 = \left\|\sum_{k\in S^{(b)}} \frac{\sqrt{\widehat{\boldsymbol{D}}_{kk}}}{vol(\mathcal{G}^{(b)})}\alpha_k\right\|_2, \tag{15}$$

$$C_2 = \sum_{k\in\mathcal{V}} \|\alpha_k\|_2. \tag{16}$$

In simpler terms, Theorem 4.3 states that with social stratification and expansion, the expected LP score $\mathbb{E}\left[f_{LP}\left(\boldsymbol{s}_i^{(L)}, \boldsymbol{s}_j^{(L)}\right)\right] \propto \sqrt{\widehat{\boldsymbol{D}}_{ii}\widehat{\boldsymbol{D}}_{jj}}$ approximately when i, j belong to the same social group. This is because, as explained before Theorem 4.3, $\zeta_s \approxeq 0$, so the RHS of the bound is $\approxeq 0$. This demonstrates that in LP, GCNs with a symmetric normalized graph filter have a within-group PA bias. If Φ_s positively influences the formation of links over time, this PA bias can drive "rich get richer" dynamics within social groups (Stoica et al., 2018). As shown in Figure 1 and §4.2, such "rich get richer" dynamics can engender group unfairness when nodes' degrees are statistically associated with their group membership (§4.2). An association between node degree and group membership

depends on group size and homophily; in particular, when a group has many nodes and intra-links (i.e., is homophilous), there may be more nodes with a high within-group degree. Beyond fairness, Theorem 4.3 reveals that GCNs do not align with theories that *social rank* influences link formation, i.e., the likelihood of a link forming between nodes is proportional to their degree *difference* (Gu et al., 2018).

4.2. Within-Group Fairness

We further investigate the fairness implications of the PA bias of Φ_s in LP. We first introduce an additional set of social groups. Suppose that $\mathcal{V}$ can also be partitioned into D disjoint social groups $\{T^{(d)}\}_{d\in[D]}$; then, we can consider intersections of $\{S^{(b)}\}_{b\in[B]}$ and $\{T^{(d)}\}_{d\in[D]}$. For example, revisiting Figure 1, S may correspond to academic discipline (e.g., CS, EDU) and T may correspond to gender (e.g., men, women). For simplicity, we let $D = 2$. We measure the unfairness $\Delta^{(b)} : \mathbb{R}^{d'} \times \mathbb{R}^{d'} \to \mathbb{R}$ of LP for group b as:

$$\Delta^{(b)}\left(\boldsymbol{h}_i^{(L)}, \boldsymbol{h}_j^{(L)}\right) := \tag{17}$$

$$\left| \mathop{\mathbb{E}}_{i,j\sim U((S^{(b)}\cap T^{(1)})\times S^{(b)})} f_{LP}\left(\boldsymbol{h}_i^{(L)}, \boldsymbol{h}_j^{(L)}\right) \right. \tag{18}$$

$$\left. - \mathop{\mathbb{E}}_{i,j\sim U((S^{(b)}\cap T^{(2)})\times S^{(b)})} f_{LP}\left(\boldsymbol{h}_i^{(L)}, \boldsymbol{h}_j^{(L)}\right) \right|, \tag{19}$$

where $U(\cdot)$ is a discrete uniform distribution over the input set. $\Delta^{(b)}$ quantifies disparities in GCN LP scores within $S^{(b)}$ (with respect to $T^{(1)}$ and $T^{(2)}$). In other words, $\Delta^{(b)}$ measures differences in how GCNs allocate LP scores across subgroups, i.e., are links with nodes in one subgroup predicted at a higher rate than links with nodes in the other subgroup? Our metric is motivated by how GNN link predictions influence real-world link formation (e.g., GNN-based recommender systems use LP scores to rank suggested social connections), which has consequences for degree and power disparities. Based on Theorem 4.3 and §B.1, when $\zeta_s \approxeq 0$, we can estimate $\Delta^{(b)}\left(\boldsymbol{s}_i^{(L)}, \boldsymbol{s}_j^{(L)}\right)$ as:

$$\widehat{\Delta}^{(b)}\left(\boldsymbol{s}_i^{(L)}, \boldsymbol{s}_j^{(L)}\right) \tag{20}$$

$$= \frac{\overline{\rho}_s^2(b)}{|S^{(b)}|} \left\| \sum_{k\in S^{(b)}} \frac{\sqrt{\widehat{\boldsymbol{D}}_{kk}}}{\text{vol}(\mathcal{G}^{(b)})} \alpha_k \right\|_2^2 \left| \sum_{j\in S^{(b)}} \sqrt{\widehat{\boldsymbol{D}}_{jj}} \times \right. \tag{21}$$

$$\underbrace{\left. \left(\mathop{\mathbb{E}}_{i\sim U(S^{(b)}\cap T^{(1)})} \sqrt{\widehat{\boldsymbol{D}}_{ii}} - \mathop{\mathbb{E}}_{i\sim U(S^{(b)}\cap T^{(2)})} \sqrt{\widehat{\boldsymbol{D}}_{ii}} \right) \right|}_{\text{degree disparity}} \tag{22}$$

This suggests that a large disparity in the degree of nodes in $S^{(b)}\cap T^{(1)}$ vs. $S^{(b)}\cap T^{(2)}$ can greatly increase the unfairness $\Delta^{(b)}$ of Φ_s LP. For example, in Figure 1, the large degree disparity within CS (between men and women) entails that a GCN collaboration recommender system applied to the network will have a large $\Delta^{(b)}$. We empirically validate these fairness implications on diverse network datasets in §6.2. While we consider pre-activation LP scores in Eqn. 17 (in line with prior work, e.g., Li et al. (2021)), we consider post-sigmoid scores $\sigma\left(f_{LP}\left(\boldsymbol{h}_i^{(L)}, \boldsymbol{h}_j^{(L)}\right)\right)$ (where σ is the sigmoid function) in §6.2 and §6.3, as this simulates how LP scores may be processed in practice.

Ultimately, within-group unfairness is characteristic of all GNN link prediction methods that: (1) predict scores for links with magnitudes that are positively associated with the degrees of their incident nodes, and (2) are applied to graphs where within-group membership is associated with node degree.

4.3. Random Walk Normalized Filter

We now follow similar steps as with Φ_s to understand how degree bias affects LP scores for Φ_r. We redefine $\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{A}$, $\widehat{\boldsymbol{P}} = \widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{A}}$, and the remaining notation from §4.1 accordingly for the random walk setting.

Theorem 4.4. *Let* $\zeta_r = \max_{u,v\in\mathcal{V}} \sqrt{\frac{\widehat{\boldsymbol{D}}_{vv}}{\widehat{\boldsymbol{D}}_{uu}}} \left(\lambda^{(b)}\right)^L + \sum_{l=1}^{L} \binom{L}{l} \left\|\Xi^{(0)}\right\|_{op}^{l} \left\|\widehat{\boldsymbol{P}}\right\|_{op}^{L-l}$. *Furthermore, for nodes* $i, j \in S^{(b)}$, *assume that* $\rho_r(i) = \rho_r(j) = \overline{\rho}_r(b)$. *Combining Lemmas 4.1 and A.1:*

$$\left| \mathbb{E}\left[f_{LP}\left(\boldsymbol{r}_i^{(L)}, \boldsymbol{r}_j^{(L)}\right)\right] - C_0 \right| \tag{23}$$

$$\leq \zeta_r \overline{\rho}_r^2(b) C_1 C_2 + \zeta_r^2 \overline{\rho}_r^2(b) C_2^2, \tag{24}$$

where: (25)

$$C_0 = \overline{\rho}_r^2(b) C_1^2, \tag{26}$$

$$C_1 = \left\| \sum_{k\in S^{(b)}} \frac{\widehat{\boldsymbol{D}}_{kk}}{vol(\mathcal{G}^{(b)})} \beta_k \right\|, \tag{27}$$

$$C_2 = \sum_{k\in\mathcal{V}} \|\beta_k\|_2. \tag{28}$$

In other words, if $\zeta_r \approxeq 0$, $\mathbb{E}\left[f_{LP}\left(\boldsymbol{r}_i^{(L)}, \boldsymbol{r}_j^{(L)}\right)\right]$ is approximately constant when i, j belong to the same social group. Based on Theorem 4.4 and §B.2, we can estimate $\Delta^{(b)}\left(\boldsymbol{s}_i^{(L)}, \boldsymbol{s}_j^{(L)}\right)$ as $\widehat{\Delta}^{(b)}\left(\boldsymbol{s}_i^{(L)}, \boldsymbol{s}_j^{(L)}\right) = 0$. Theoretically, this would suggest that a large disparity in the degree of nodes in $S^{(b)}\cap T^{(1)}$ vs. $S^{(b)}\cap T^{(2)}$ does not increase the unfairness $\Delta^{(b)}$ of Φ_r LP. However, we find empirically that this is not the case (§6.1). Even so, we include theoretical results for the random walk filter to be more comprehensive with respect to filter choice, as well as be upfront about the limitations of our analysis in this case. We also seek to provide an example of how to apply our analysis to other

filters, for researchers who would like to build on it in the future. For example, findings for the random walk filter could be relevant to the GAT filter (Veličković et al., 2018), which is also a row-stochastic matrix.

In summary, in §4, we build on prior analysis techniques for random walks and GNNs. At a high level, we: (1) simplify the GCN architecture to be a linear function by truncating its Taylor expansion and considering node representations in expectation; (2) analyze the convergence of node representations via a spectral analysis of the convergence of short random walks within subgraphs (corresponding to social groups); and (3) use norm inequalities to estimate link prediction scores. Our analysis comprises numerous novel elements including:

1. Analyzing the convergence of random walks within subgraphs, which requires accounting for the rate at which probability mass escapes from the subgraphs. In contrast, random walk results in the literature usually concern the convergence of random walks over an entire graph.
2. Uncovering properties of short random walks on graphs, since most GNNs are shallow. In contrast, random walk results in the literature often concern the stationary distribution of random walks.
3. Concretely relating theoretical properties of random walks to the fairness of GCN link prediction.

5. Fairness Regularizer

We propose a simple training-time solution to alleviate within-group LP unfairness regardless of graph filter type and GNN architecture. In particular, we can add a fairness regularization term $\mathcal{L}_{\text{fair}}$ to our original GNN training loss (Kamishima et al., 2011):

$$\mathcal{L}_{\text{new}} = \mathcal{L}_{\text{orig}} + \lambda_{\text{fair}}\mathcal{L}_{\text{fair}} = \mathcal{L}_{\text{orig}} + \frac{\lambda_{\text{fair}}}{B}\sum_{b\in[B]}\Delta^{(b)}, \quad (29)$$

where λ_{fair} is a tunable hyperparameter that for higher values, pushes the GNN to learn fairer parameters. With our fairness strategy, we empirically observe a significant decrease in the average unfairness across groups $\frac{1}{B}\sum_{b\in[B]}\Delta^{(b)}$ without a severe drop in LP performance for GCN (§6.3).

6. Experiments

In this section, we empirically validate our theoretical analysis (§6.1) and the within-group fairness implications of GCN's LP PA bias (§6.2) on diverse real-world network datasets of varying size. We further find that our simple training-time strategy to alleviate unfairness is effective on citation, online social, and credit networks (§6.3). We release our code and data in our GitHub repository[1]. We present experimental results with 4-layer GCN encoders and a Hadamard product with MLP LP score function in §G, with similar conclusions.

6.1. Validating Theoretical Analysis

We validate our theoretical analysis on 10 real-world network datasets (e.g., citation, collaboration, online social networks), which we describe in §C. Each dataset is natively intended for node classification; however, we adapt the datasets for LP, treating the connected components within the node classes as the social groups $S^{(b)}$. This design choice is reasonable, as in all the datasets, the classes naturally correspond to socially-relevant groupings of the nodes, or proxies thereof (e.g., in the LastFMAsia dataset, the classes are the home countries of users). Because we adopt the class labels for each dataset as the social group labels, the social groups are largely homophilic; this aligns with our assumptions when interpreting Theorems 4.3 and 4.4 that social groups are stratified in networks.

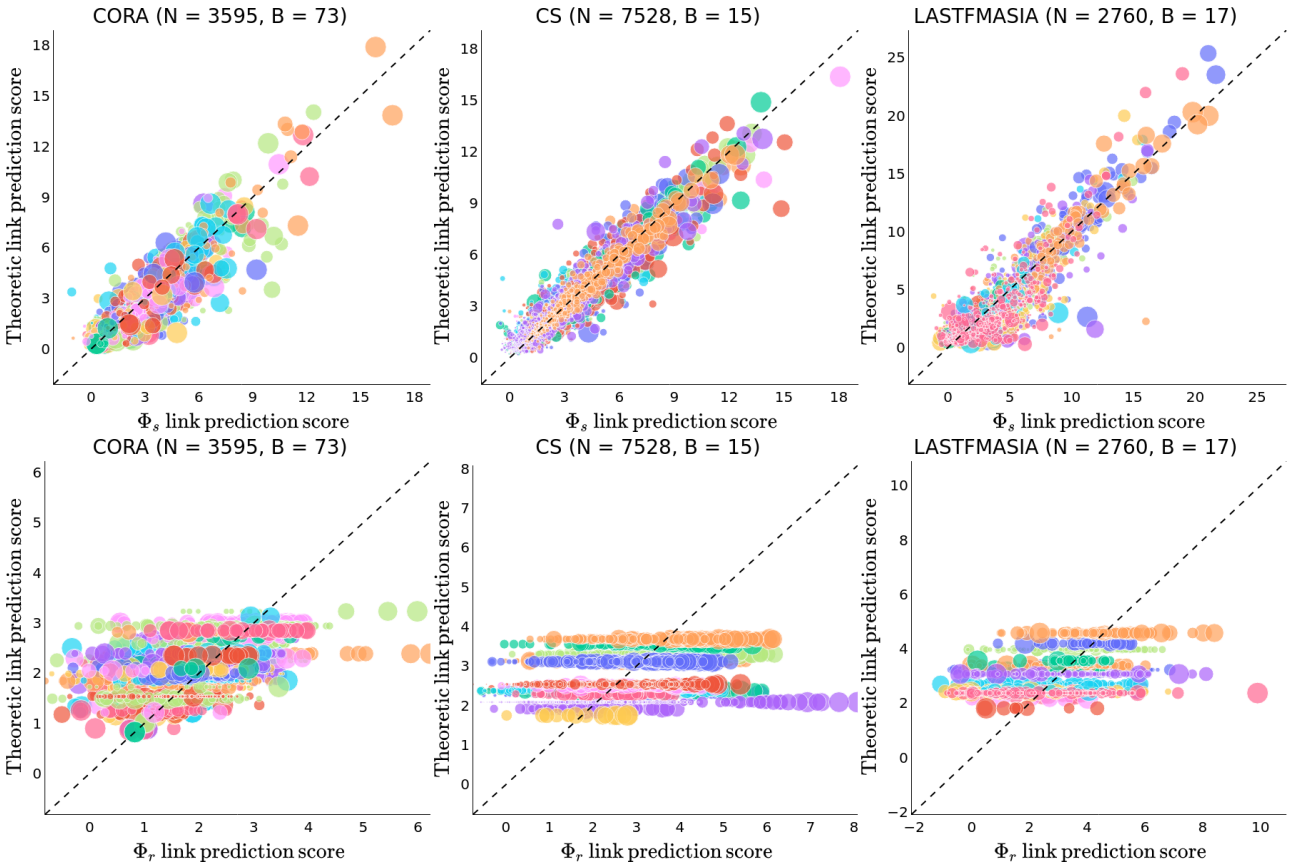
We train GCN encoders Φ_s and Φ_r for LP over 10 random seeds (cf. §E for more details). In Figure 2, we plot the theoretic[2] LP score that we derive in §4 against the GCN LP score *for pairs of test nodes belonging to the same social group* (including positive and negative links). In particular, for Φ_s, the theoretic LP score is $\overline{\rho}_s^2(b)\sqrt{\widehat{\boldsymbol{D}}_{ii}\widehat{\boldsymbol{D}}_{jj}}\left\|\sum_{k\in S^{(b)}}\frac{\sqrt{\widehat{\boldsymbol{D}}_{kk}}}{\text{vol}(\mathcal{G}^{(b)})}\alpha_k\right\|_2^2$ and the GCN LP score is $f_{LP}\left(\boldsymbol{s}_i^{(L)}, \boldsymbol{s}_j^{(L)}\right)$ (cf. Theorem 4.3). In contrast, for Φ_r, the theoretic LP score is $\overline{\rho}_s^2(b)\left\|\sum_{k\in S^{(b)}}\frac{\widehat{\boldsymbol{D}}_{kk}}{\text{vol}(\mathcal{G}^{(b)})}\beta_k\right\|_2^2$ and the GCN LP score is $f_{LP}\left(\boldsymbol{r}_i^{(L)}, \boldsymbol{r}_j^{(L)}\right)$ (cf. Theorem 4.4). For all the datasets, we estimate $\overline{\rho}_s^2(b)$ and $\overline{\rho}_r^2(b)$ separately for each social group $S^{(b)}$ as the slope of the least-squares regression line (through the data from $S^{(b)}$) that predicts the GCN score as a function of the theoretic score. Hence, we do not plot any pair of test nodes that is the only pair in $S^{(b)}$, as it is not possible to estimate $\overline{\rho}_s^2(b)$. Further, the test AUC is consistently high, indicating that the GCNs are well-trained. The large range of each color in the plots indicates a diversity of LP scores within each social group.

We visually observe that the theoretic LP scores are strong predictors of the Φ_s scores for each dataset, validating our theoretical analysis. This strength is further confirmed by

[1] https://github.com/ArjunSubramonian/link_bias_amplification

[2] While our theoretic scores resulted from our theoretical analysis in §4, we reiterate that our results in §4 rely on the assumptions that we state and the theoretic score is not a ground-truth value.

[3] Normalized by the sample range of the GCN LP scores. Values fall between 0 and 1.



	NRMSE (↓)	PCC (↑)	Φ_s Test AUC (↑)
CORA	0.038 ± 0.006	0.884 ± 0.008	0.927 ± 0.008
CITESEER	0.080 ± 0.005	0.806 ± 0.007	0.943 ± 0.007
DBLP	0.026 ± 0.002	0.820 ± 0.014	0.948 ± 0.001
PUBMED	0.061 ± 0.008	0.774 ± 0.018	0.927 ± 0.010
CS	0.036 ± 0.006	0.917 ± 0.019	0.932 ± 0.008
PHYSICS	0.042 ± 0.003	0.822 ± 0.021	0.946 ± 0.003
LASTFMASIA	0.064 ± 0.003	0.889 ± 0.004	0.962 ± 0.001
DE	0.025 ± 0.003	0.795 ± 0.043	0.913 ± 0.003
EN	0.041 ± 0.002	0.542 ± 0.013	0.876 ± 0.003
FR	0.030 ± 0.002	0.743 ± 0.026	0.910 ± 0.005

	NRMSE (↓)	PCC (↑)	Φ_r Test AUC (↑)
CORA	0.101 ± 0.029	0.553 ± 0.024	0.942 ± 0.005
CITESEER	0.170 ± 0.016	0.363 ± 0.028	0.934 ± 0.003
DBLP	0.157 ± 0.012	0.235 ± 0.022	0.942 ± 0.002
PUBMED	0.155 ± 0.013	0.079 ± 0.029	0.896 ± 0.011
CS	0.101 ± 0.027	0.447 ± 0.070	0.939 ± 0.003
PHYSICS	0.107 ± 0.027	0.264 ± 0.038	0.951 ± 0.004
LASTFMASIA	0.123 ± 0.016	0.409 ± 0.017	0.949 ± 0.001
DE	0.024 ± 0.004	0.074 ± 0.016	0.862 ± 0.003
EN	0.065 ± 0.006	0.012 ± 0.005	0.850 ± 0.002
FR	0.028 ± 0.006	0.006 ± 0.003	0.865 ± 0.004

Figure 2: The plots display the theoretic vs. GCN LP scores for the Cora, CS, and LastFMAsia datasets over 10 random seeds. (We include the plots for the remaining datasets in §F.) The **top row** of plots corresponds to Φ_s, the **bottom row** to Φ_r. In the plots, each circle corresponds to a single pair of test nodes (between which we are predicting a link). The center of each circle represents the mean of the theoretic and GCN scores and its area captures the range of scores. The color of each circle indicates the social group to which the node pair belongs. The plots include: (1) the total number of test node pairs N; (2) the number of social groups B; (3) the dashed line of equality for easy comparison of the theoretic and GCN scores. For all the datasets, the tables display: (1) the mean/standard deviation of the GCN test AUC on LP; and (2) the mean/standard deviation of the range-normalized[3]root-mean-square deviation (NRMSE) (Otto, 2019) and Pearson correlation coefficient (PCC) (Freedman et al., 2007) of the theoretic LP scores as predictors of the GCN scores. The **left** table corresponds to Φ_s, the **right** to Φ_r.

the generally low NRMSE and high PCC (except for the EN dataset). However, we observe a few cases in which our theoretical analysis does not line up with our experiments:

1. Our theoretical analysis predicts that the LP score between two nodes i, j that belong to the same social group $S^{(b)}$ will always be non-negative; however, Φ_s can predict negative scores for pairs of nodes in the same social

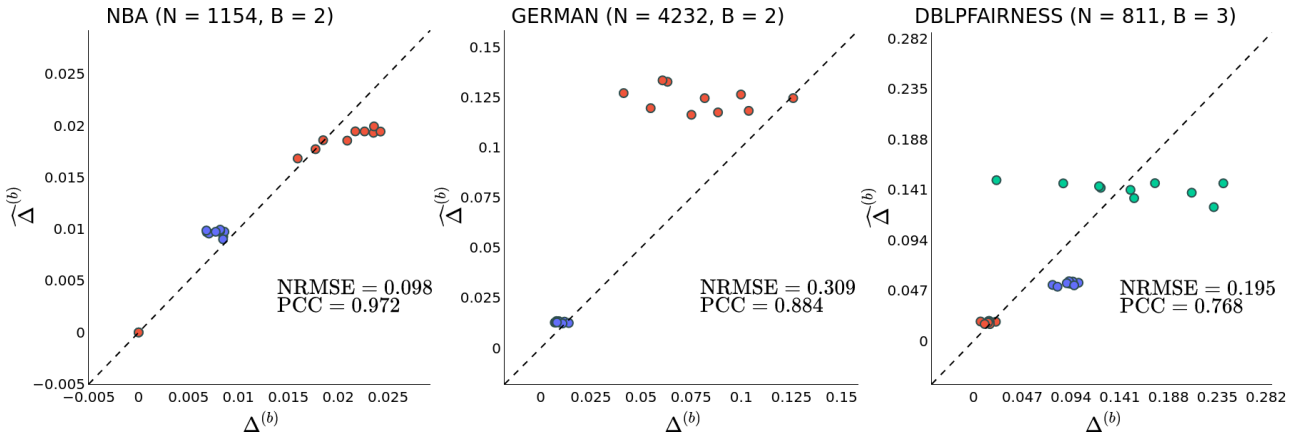


Figure 3: The plots display $\widehat{\Delta}^{(b)}$ vs. $\Delta^{(b)}$ for Φ_s for the NBA, German, and DBLP-Fairness datasets over all $b \in [B]$ and 10 random seeds. Each point corresponds to a different random seed, and the color of the point corresponds to the social group $S^{(b)}$. We compute $\widehat{\Delta}^{(b)}$ and $\Delta^{(b)}$ post-sigmoid using only the LP scores over the sampled (positive and negative) test edges. The plots display the NRMSE and PCC of $\widehat{\Delta}^{(b)}$ as a predictor of $\Delta^{(b)}$.

group. In this case, it appears that Φ_s relies more on the dissimilarity of (transformed) features than node degree.

2. For many network datasets (especially from the citation and online social domains), there exist node pairs (near the origin) for which the theoretic LP score underestimates the Φ_s score. Upon further analysis (cf. Appendix H), we find that the theoretic score is less predictive of the Φ_s score for nodes i, j when the product of their degrees (i.e., their PA score) or similarity of their features is relatively low.
3. It appears that the theoretic LP score tends to poorly estimate the Φ_s score when the Φ_s score is relatively high; this suggests that Φ_s may conservatively rely more on the (dis)similarity of node features than node degree when the degree is large.

We do not observe that the theoretic LP scores are strong predictors of the Φ_r scores, although there is still a moderate association between these variables. This could be because the error bound for the theoretic scores for Φ_r, unlike for Φ_s, has an extra dependence $\max_{u,v \in \mathcal{V}} \sqrt{\frac{\widetilde{\boldsymbol{D}}_{vv}}{\widetilde{\boldsymbol{D}}_{uu}}}$ on the degrees of the incident nodes (cf. ζ_r in Theorem 4.4). In contrast, the error bound for the theoretic scores for Φ_s (cf. ζ_s in Theorem 4.3) does not depend on this degree ratio. This ratio can be quite large in social networks (e.g., celebrities vs. new users in the Twitter follow network); we further confirm that this ratio is large for our datasets in §I.

6.2. Within-Group Fairness

We now empirically validate the implications of GCN's PA bias for within-group unfairness in LP. We run experiments on three network datasets: (1) the NBA social network (Dai & Wang, 2021), (2) the German credit network (Agarwal et al., 2021), and (3) a new DBLP-Fairness citation network that we construct. We describe these datasets in §D, including $\{S^{(b)}\}_{b \in [B]}$ and $\{T^{(d)}\}_{d \in [D]}$.

We train 2-layer GCN encoders Φ_s for LP (cf. §E). In Figure 3, for all the datasets, we plot $\widehat{\Delta}^{(b)}$ vs. $\Delta^{(b)}$ (cf. Eqns. 17, 22) for each $b \in [B]$. We qualitatively and quantitatively observe that $\widehat{\Delta}^{(b)}$ is moderately predictive of $\Delta^{(b)}$ for each dataset. This confirms our theoretical intuition (§4.2) that a large disparity in the degree of nodes in $S^{(b)} \cap T^{(1)}$ vs. $S^{(b)} \cap T^{(2)}$ can greatly increase the unfairness $\Delta^{(b)}$ of Φ_s LP; such unfairness can amplify degree disparities, worsening power imbalances in the network. Many points deviate from the line of equality; these deviations can be explained by the reasons in §6.1 and the compounding of errors.

6.3. Fairness Regularizer

We evaluate our solution to alleviate LP unfairness (§4.2). In particular, we add our fairness regularization term $\mathcal{L}_{\text{fair}}$ to the original training loss for the 2-layer Φ_s and Φ_r encoders. During each training epoch, we compute $\Delta^{(b)}$ post-sigmoid using only the LP scores over the sampled (positive and negative) training edges. In Table 1, we summarize the link prediction fairness $\left(\frac{1}{B}\sum_{b \in [B]} \Delta^{(b)}\right)$ and performance (test AUC) for the NBA, German, and DBLP-Fairness datasets with various settings of λ_{fair}.

For both graph filter types, we generally observe a significant decrease in $\frac{1}{B}\sum_{b \in [B]} \Delta^{(b)}$ (without a severe drop in test AUC) for $\lambda_{\text{fair}} > 0.0$ over $\lambda_{\text{fair}} = 0.0$ (with the exception of Φ_r for German); however, the varying magnitudes by which $\frac{1}{B}\sum_{b \in [B]} \Delta^{(b)}$ decreases across the datasets suggests that λ_{fair} may need to be tuned per dataset. As expected, we mostly observe a tradeoff between $\frac{1}{B}\sum_{b \in [B]} \Delta^{(b)}$ and

Table 1: $\frac{1}{B}\sum_{b\in[B]}\Delta^{(b)}$ and the test AUC for the NBA, German, and DBLP-Fairness datasets with various settings of λ_{fair}. The **left** table corresponds to Φ_s, and the **right** to Φ_r.

	λ_{fair}	$\frac{1}{B}\sum_{b\in[B]}\Delta^{(b)}$ (↓)	Φ_s **Test AUC** (↑)
NBA	4.0	0.000 ± 0.001	0.753 ± 0.002
NBA	2.0	0.004 ± 0.003	0.752 ± 0.003
NBA	1.0	0.007 ± 0.004	0.752 ± 0.003
NBA	0.0	0.013 ± 0.005	0.752 ± 0.003
DBLPFAIRNESS	4.0	0.072 ± 0.018	0.741 ± 0.008
DBLPFAIRNESS	2.0	0.095 ± 0.025	0.756 ± 0.007
DBLPFAIRNESS	1.0	0.110 ± 0.033	0.770 ± 0.010
DBLPFAIRNESS	0.0	0.145 ± 0.020	0.778 ± 0.007
GERMAN	4.0	0.012 ± 0.006	0.876 ± 0.017
GERMAN	2.0	0.028 ± 0.017	0.889 ± 0.017
GERMAN	1.0	0.038 ± 0.016	0.897 ± 0.014
GERMAN	0.0	0.045 ± 0.013	0.912 ± 0.009

	λ_{fair}	$\frac{1}{B}\sum_{b\in[B]}\Delta^{(b)}$ (↓)	Φ_r **Test AUC** (↑)
NBA	4.0	0.000 ± 0.000	0.585 ± 0.030
NBA	2.0	0.000 ± 0.000	0.584 ± 0.032
NBA	1.0	0.000 ± 0.000	0.581 ± 0.034
NBA	0.0	0.000 ± 0.000	0.583 ± 0.034
DBLPFAIRNESS	4.0	0.053 ± 0.015	0.715 ± 0.010
DBLPFAIRNESS	2.0	0.060 ± 0.016	0.731 ± 0.009
DBLPFAIRNESS	1.0	0.065 ± 0.022	0.746 ± 0.009
DBLPFAIRNESS	0.0	0.090 ± 0.028	0.758 ± 0.011
GERMAN	4.0	0.029 ± 0.011	0.830 ± 0.024
GERMAN	2.0	0.031 ± 0.019	0.843 ± 0.027
GERMAN	1.0	0.019 ± 0.012	0.864 ± 0.020
GERMAN	0.0	0.015 ± 0.005	0.883 ± 0.009

the test AUC as λ_{fair} increases. Our experiments reveal that, regardless of graph filter type, even simple regularization approaches can alleviate this new form of unfairness. As this form of unfairness has not been previously explored, we have no baselines.

Our fairness regularizer can be easily integrated into model training, does not require significant additional computation, and directly optimizes for LP fairness. The time complexity of calculating the regularization term is $\mathcal{O}\left(\sum_{b=1}^{B}|S^{(b)}\cap T^{(1)}|\cdot|S^{(b)}|+|S^{(b)}\cap T^{(2)}|\cdot|S^{(b)}|\right)$, as we have already computed the LP scores for the cross-entropy loss term and simply need to sum them appropriately with respect to the groups and subgroups. Furthermore, the time complexity of computing gradients for the regularization term is on the same order as backpropagation for the cross-entropy loss term.

However, our fairness regularizer is not applicable in settings where model parameters cannot be retrained or fine-tuned. Hence, we encourage future research to also explore post-processing fairness strategies. For example, for Φ_s models, based on our theory (cf. Theorem 4.3), for each pair of nodes i, j, we can decay the influence of GCN's PA bias by scaling (pre-activation) LP scores by $\left(\sqrt{\widehat{\boldsymbol{D}}_{ii}\widehat{\boldsymbol{D}}_{jj}}\right)^{-\alpha}$, where $0 < \alpha < 1$ is a hyperparameter that can be tuned to achieve a desirable balance between $\frac{1}{B}\sum_{b\in[B]}\Delta^{(b)}$ and the test AUC.

Empirical evaluation of our fairness regularizer using existing LP fairness metrics, such as statistical parity and equal opportunity dyadic fairness (Li et al., 2021), or equal opportunity degree bias (Wang & Derr, 2022), is beyond the scope of our paper given that our algorithm and metric are designed to handle a different form of unfairness. For example, inter-group and intra-group links can be predicted at the same rate or with the same accuracy, but these links can be exclusively with high-degree nodes, thereby marginalizing low-degree nodes (cf. §J). Similarly, even if we consistently predict links with the same accuracy across nodes with different degrees, high-degree nodes can still receive higher LP scores than low-degree nodes (cf. §K).

7. Conclusion

We theoretically and empirically show that GCNs can have a PA bias in LP. We analyze how this bias can engender within-group unfairness, and amplify degree and power imbalances in networks. We further propose a simple training-time strategy to alleviate this unfairness. We encourage future work to: (1) explore PA bias in other GNN architectures and directed and heterophilic networks, (2) characterize the "rich get richer" evolution of networks affected by GCN's PA bias, and (3) propose pre-processing and post-processing strategies for within-group LP unfairness.

Because this unfairness is at the level of dyads, we would like to explore new forms of unfairness that occur at the level of higher-order structures (e.g., prediction disparities between important coalitions of nodes). Moreover, node degree is a local property, and it would be valuable to theoretically and empirically relate higher-order graph properties (e.g., local clustering coefficient, different measures of centrality) to unfairness.

Acknowledgements

We would like to thank the anonymous reviewers for their feedback on this work. This work was partially supported by NSF 2211557, NSF 1937599, NSF 2119643, NSF 2303037, NSF 2312501, NASA, SRC JUMP 2.0 Center, Amazon Research Awards, and Snapchat Gifts.

Impact Statement

Our paper seeks to uncover and combat discrimination, bias, and unfairness in GNNs. Throughout, we tie our analysis back to issues of disparity and power, towards advancing justice in graph learning. While we propose a strategy to alleviate LP unfairness, we emphasize that it is not a 'silver bullet' solution; we encourage graph learning practitioners to adopt a sociotechnical approach to fairness and continually adapt their algorithms, datasets, and metrics in response to the everchanging landscape of inequality and power. Furthermore, the fairness of GCN LP should not sidestep concerns about GCN LP being used *at all* in certain scenarios.

Some datasets that we use contain protected attribute information (detailed in §D). We avoid using datasets that enable carceral technology (e.g., Recidivism (Agarwal et al., 2021)). We release our code and data with an MIT license.

For transparency, we do our best to discuss limitations throughout the paper. For each lemma and theorem (§4), our assumptions are clearly explained and justified either before or in the statement thereof, and we include complete proofs of our theoretical claims in §A and §B.

For reproducibility, we provide all our code and data (including the raw DBLP-Fairness dataset) in our GitHub repository, along with a README. We detail our data processing steps in §D.3. Furthermore, our experiments (§6) are run with 10 random seeds and errors are reported. We provide model implementation details in §E.

References

Agarwal, C., Lakkaraju, H., and Zitnik, M. Towards a unified framework for fair and stable graph representation learning. In *Conference on Uncertainty in Artificial Intelligence*, 2021.

Barabási, A.-L. and Albert, R. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999. doi: 10.1126/science.286.5439.509. URL https://www.science.org/doi/abs/10.1126/science.286.5439.509.

Bashardoust, A., Friedler, S. A., Scheidegger, C. E., Sullivan, B. D., and Venkatasubramanian, S. Reducing access disparities in networks using edge augmentation. *ArXiv*, abs/2209.07616, 2022.

Bojchevski, A. and Günnemann, S. Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1ZdKJ-0W.

Boyd, D., Levy, K., and Marwick, A. The networked nature of algorithmic discrimination. *Data and Discrimination: Collected Essays. Open Technology Institute*, 2014.

Chamberlain, B. P., Shirobokov, S., Rossi, E., Frasca, F., Markovich, T., Hammerla, N. Y., Bronstein, M. M., and Hansmire, M. Graph neural networks for link prediction with subgraph sketching. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=m1oqEOAozQU.

Collins, P. H. and Bilge, S. *Intersectionality*. John Wiley & Sons, 2020.

Current, S., He, Y., Gurukar, S., and Parthasarathy, S. FairEGM: Fair link prediction and recommendation via emulated graph modification. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, oct 2022. doi: 10.1145/3551624.3555287. URL https://doi.org/10.1145%2F3551624.3555287.

Dai, E. and Wang, S. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, pp. 680–688, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450382977. doi: 10.1145/3437963.3441752. URL https://doi.org/10.1145/3437963.3441752.

Fan, W., Ma, Y., Li, Q., He, Y., Zhao, Y. E., Tang, J., and Yin, D. Graph neural networks for social recommendation. *The World Wide Web Conference*, 2019.

Fey, M. link_pred.py, 2019. URL https://github.com/pyg-team/pytorch_geometric/blob/master/examples/link_pred.py.

Fey, M. and Lenssen, J. E. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Fish, B., Bashardoust, A., Boyd, D., Friedler, S., Scheidegger, C., and Venkatasubramanian, S. Gaps in information access in social networks? In *The World Wide Web Conference*, WWW '19, pp. 480–490, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313680. URL https://doi.org/10.1145/3308558.3313680.

Foulds, J. R., Islam, R., Keya, K. N., and Pan, S. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pp. 1918–1921, 2020. doi: 10.1109/ICDE48307.2020.00203.

Freedman, D., Pisani, R., and Purves, R. *Statistics: Fourth International Student Edition*. Emersion: Emergent Village Resources for Communities of Faith Series. W.W. Norton & Company, 2007. ISBN 9780393930436.

Ghosh, A., Genuit, L., and Reagan, M. Characterizing intersectional group fairness with worst-case comparisons. In *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*, pp. 22–34. PMLR, 2021.

Giovanni, F. D., Rowbottom, J., Chamberlain, B. P., Markovich, T., and Bronstein, M. M. Understanding convolution on graphs via energies. 2023. URL https://openreview.net/forum?id=v5ew3FPTgb.

Gu, Y., Sun, Y., Li, Y., and Yang, Y. Rare: Social rank regulated large-scale network embedding. *Proceedings of the 2018 World Wide Web Conference*, 2018.

Hofstra, B., Corten, R., van Tubergen, F., and Ellison, N. B. Sources of segregation in social networks: A novel approach using facebook. *American Sociological Review*, 82(3):625–656, 2017. doi: 10.1177/0003122417705656. URL https://doi.org/10.1177/0003122417705656.

Kamishima, T., Akaho, S., and Sakuma, J. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 643–650, 2011. doi: 10.1109/ICDMW.2011.83.

Kang, J., Zhu, Y., Xia, Y., Luo, J., and Tong, H. Rawlsgcn: Towards rawlsian difference principle on graph convolutional network. In *Proceedings of the ACM Web Conference 2022*, pp. 1214–1225, 2022.

Kasy, M. and Abebe, R. Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 576–586, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445919. URL https://doi.org/10.1145/3442188.3445919.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, 2017.

Keriven, N. Not too little, not too much: a theoretical analysis of graph (over)smoothing. *ArXiv*, abs/2205.12156, 2022.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Leskovec, J., Lang, K. J., Dasgupta, A., and Mahoney, M. W. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6:123–29, 2008.

Li, P., Wang, Y., Wang, H., and Leskovec, J. Distance encoding: Design provably more powerful neural networks for graph representation learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Li, P., Wang, Y., Zhao, H., Hong, P., and Liu, H. On dyadic fairness: Exploring and mitigating bias in graph connections. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=xgGS6PmzNq6.

Li, Y., Wang, X., Ning, Y., and Wang, H. Fairlp: Towards fair link prediction on social network graphs. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):628–639, May 2022. doi: 10.1609/icwsm.v16i1.19321. URL https://ojs.aaai.org/index.php/ICWSM/article/view/19321.

Liu, Z., Nguyen, T.-K., and Fang, Y. Tail-gnn: Tail-node graph neural networks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pp. 1109–1119, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467276. URL https://doi.org/10.1145/3447548.3467276.

Liu, Z., Nguyen, T.-K., and Fang, Y. On generalized degree fairness in graph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):4525–4533, Jun. 2023. doi: 10.1609/aaai.v37i4.25574. URL https://ojs.aaai.org/index.php/AAAI/article/view/25574.

Lovász, L. M. Random walks on graphs: A survey. 2001.

Malliaros, F. D. and Megalooikonomou, V. Expansion properties of large social graphs. In *DASFAA Workshops*, 2011.

Nakkiran, P., Kaplun, G., Kalimeris, D., Yang, T., Edelman, B. L., Zhang, F., and Barak, B. Sgd on neural networks learns functions of increasing complexity, 2019.

Oono, K. and Suzuki, T. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1ldO2EFPr.

Otto, S. How to normalize the rmse [blog post], 2019. URL https://www.marinedatascience.co/blog/2019/01/07/normalizing-the-rmse/.

Ovalle, A., Subramonian, A., Gautam, V., Gee, G., and Chang, K.-W. Factoring the matrix of domination: A critical review and reimagination of intersectionality in ai fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pp. 496–511, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702310. doi: 10.1145/3600211.3604705. URL https://doi.org/10.1145/3600211.3604705.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Rozemberczki, B. and Sarkar, R. Characteristic functions on graphs: Birds of a feather, from statistical descriptors to parametric models. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pp. 1325–1334, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368599. doi: 10.1145/3340531.3411866. URL https://doi.org/10.1145/3340531.3411866.

Rozemberczki, B., Allen, C., and Sarkar, R. Multi-Scale attributed node embedding. *Journal of Complex Networks*, 9(2):cnab014, 05 2021. ISSN 2051-1329. doi: 10.1093/comnet/cnab014. URL https://doi.org/10.1093/comnet/cnab014.

Sankar, A., Liu, Y., Yu, J., and Shah, N. Graph neural networks for friend ranking in large-scale social platforms. *Proceedings of the Web Conference 2021*, 2021.

Shchur, O., Mumme, M., Bojchevski, A., and Günnemann, S. Pitfalls of graph neural network evaluation. *ArXiv*, abs/1811.05868, 2018.

Shomer, H., Jin, W., Wang, W., and Tang, J. Toward degree bias in embedding-based knowledge graph completion. In *Proceedings of the ACM Web Conference 2023*, WWW '23, pp. 705–715, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394161. doi: 10.1145/3543507.3583544. URL https://doi.org/10.1145/3543507.3583544.

Stoica, A.-A., Riederer, C., and Chaintreau, A. Algorithmic glass ceiling in social networks: The effects of social recommendations on network diversity. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pp. 923–932, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. ISBN 9781450356398. doi: 10.1145/3178876.3186140. URL https://doi.org/10.1145/3178876.3186140.

Stoica, A.-A., Han, J. X., and Chaintreau, A. Seeding network influence in biased networks and the benefits of diversity. In *Proceedings of The Web Conference 2020*, WWW '20, pp. 2089–2098, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370233. doi: 10.1145/3366423.3380275. URL https://doi.org/10.1145/3366423.3380275.

Subramonian, A., Chang, K.-W., and Sun, Y. On the discrimination risk of mean aggregation feature imputation in graphs. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 32957–32973. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/d4c2f25bf0c33065b7d4fb9be2a9add1-Paper-Conference.pdf.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z. Arnetminer: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pp. 990–998, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581934. doi: 10.1145/1401890.1402008. URL https://doi.org/10.1145/1401890.1402008.

Tang, X., Yao, H., Sun, Y., Wang, Y., Tang, J., Aggarwal, C., Mitra, P., and Wang, S. Investigating and mitigating degree-related biases in graph convoltuional networks. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, pp. 1435–1444, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368599. doi: 10.1145/3340531.3411872. URL https://doi.org/10.1145/3340531.3411872.

Valle-Pérez, G., Camargo, C. Q., and Louis, A. A. Deep learning generalizes because the parameter-function map is biased towards simple functions, 2019.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Wang, A., Ramaswamy, V. V., and Russakovsky, O. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 336–349, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533101. URL https://doi.org/10.1145/3531146.3533101.

Wang, Y. and Derr, T. Degree-related bias in link prediction. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 757–758, 2022. doi: 10.1109/ICDMW58026.2022.00103.

Xie, Q., Zhu, Y., Huang, J., Du, P., and Nie, J.-Y. Graph neural collaborative topic model for citation recommendation. *ACM Trans. Inf. Syst.*, 40(3), nov 2021. ISSN 1046-8188. doi: 10.1145/3473973. URL https://doi.org/10.1145/3473973.

Xu, H., Xiang, L., Huang, F., Weng, Y., Xu, R., Wang, X., and Zhou, C. Grace: Graph self-distillation and completion to mitigate degree-related biases. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 2813–2824, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599368. URL https://doi.org/10.1145/3580305.3599368.

Xu, H.-R., Bu, Y., Liu, M., Zhang, C., Sun, M., Zhang, Y., Meyer, E., Salas, E., and Ding, Y. Team power dynamics and team impact: New perspectives on scientific collaboration using career age as a proxy for team power. *Journal of the Association for Information Science and Technology*, 73:1489–1505, 2021.

Xu, K., Li, C., Tian, Y., Sonobe, T., ichi Kawarabayashi, K., and Jegelka, S. Representation learning on graphs with jumping knowledge networks. In *International Conference on Machine Learning*, 2018.

Yamamoto, J. and Frachtenberg, E. Gender differences in collaboration patterns in computer science. *Publications*, 10(1), 2022. ISSN 2304-6775. doi: 10.3390/publications10010010. URL https://www.mdpi.com/2304-6775/10/1/10.

Zhang, M. and Chen, Y. Link prediction based on graph neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 5171–5181, Red Hook, NY, USA, 2018. Curran Associates Inc.

Zhang, Y., Han, J. X., Mahajan, I., Bengani, P., and Chaintreau, A. Chasm in hegemony: explaining and reproducing disparities in homophilous networks. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 5(2):1–38, 2021.

Zhao, B., Gu, Y., Forde, J. Z., and Saphra, N. One venue, two conferences: The separation of chinese and american citation networks, 2022.

Supplementary Text

A. Proofs

A.1. Proof of Lemma 4.1

Proof. Similarly to Xu et al. (2018); Tang et al. (2020), we compute the first-order partial derivatives of Φ_s and Φ_r:

$$\frac{\partial \boldsymbol{s}_i^{(L)}}{\partial \boldsymbol{x}_j} = \sum_{p \in \Psi_{i\to j}^{L+1}} \prod_{l=L}^{1} \frac{\operatorname{diag}\left(\mathbb{1}_{\boldsymbol{z}_{p^{(l)}}^{(l)}>0}\right) \boldsymbol{W}_s^{(l)}}{\sqrt{\boldsymbol{D}_{p^{(l)}p^{(l)}} \boldsymbol{D}_{p^{(l-1)}p^{(l-1)}}}}, \quad \frac{\partial \boldsymbol{r}_i^{(L)}}{\partial \boldsymbol{x}_j} = \sum_{p \in \Psi_{i\to j}^{L+1}} \prod_{l=L}^{1} \frac{\operatorname{diag}\left(\mathbb{1}_{\boldsymbol{z}_{p^{(l)}}^{(l)}>0}\right) \boldsymbol{W}_s^{(l)}}{\boldsymbol{D}_{p^{(l)}p^{(l)}}} \tag{30}$$

$$\frac{\partial \boldsymbol{s}_i^{(L)}}{\partial \boldsymbol{x}_j} = \sqrt{\frac{\boldsymbol{D}_{ii}}{\boldsymbol{D}_{jj}}} \sum_{p \in \Psi_{i\to j}^{L+1}} \prod_{l=L}^{1} \frac{\operatorname{diag}\left(\mathbb{1}_{\boldsymbol{z}_{p^{(l)}}^{(l)}>0}\right) \boldsymbol{W}_s^{(l)}}{\boldsymbol{D}_{p^{(l)}p^{(l)}}} \tag{31}$$

where $p^{(l)}$ is the l-th node on path p in the computation graph of Φ_s or Φ_r ($p^{(L)}$ is node i and $p^{(0)}$ is node j); $\Psi_{i\to j}^{\gamma}$ is the set of all γ-length random walk paths from node i to j; and $\boldsymbol{z}_{p^{(l)}}^{(l)}$ is pre-activated $\boldsymbol{s}_{p^{(l)}}^{(l)}$ or $\boldsymbol{r}_{p^{(l)}}^{(l)}$.

With our assumption that the path from node $i \to j$ in the computation graph of Φ_s is independently activated with probability $\rho_s(i)$, and similarly, $\rho_r(i)$ for Φ_r:

$$\mathbb{E}\left[\frac{\partial \boldsymbol{s}_i^{(L)}}{\partial \boldsymbol{x}_j}\right] = \left(\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}\right)_{ij}^{L} \rho_s(i) \left(\prod_{l=L}^{1} \boldsymbol{W}_s^{(l)}\right), \tag{32}$$

$$\mathbb{E}\left[\frac{\partial \boldsymbol{r}_i^{(L)}}{\partial \boldsymbol{x}_j}\right] = \left(\boldsymbol{D}^{-1}\boldsymbol{A}\right)_{ij}^{L} \rho_r(i) \left(\prod_{l=L}^{1} \boldsymbol{W}_r^{(l)}\right). \tag{33}$$

Then, recalling Eqn. 5:

$$\mathbb{E}\left[\boldsymbol{s}_i^{(L)}\right] = \sum_{j\in\mathcal{V}} \left(\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}\right)_{ij}^{L} \rho_s(i) \left(\prod_{l=L}^{1} \boldsymbol{W}_s^{(l)}\right) \boldsymbol{x}_j + \mathbf{0}, \tag{34}$$

$$\mathbb{E}\left[\boldsymbol{r}_i^{(L)}\right] = \sum_{j\in\mathcal{V}} \left(\boldsymbol{D}^{-1}\boldsymbol{A}\right)_{ij}^{L} \rho_r(i) \left(\prod_{l=L}^{1} \boldsymbol{W}_r^{(l)}\right) \boldsymbol{x}_j + \mathbf{0} \tag{35}$$

$$\mathbb{E}\left[\boldsymbol{s}_i^{(L)}\right] = \sum_{j\in\mathcal{V}} \rho_s(i) \left(\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}\right)_{ij}^{L} \alpha_j, \quad \mathbb{E}\left[\boldsymbol{r}_i^{(L)}\right] = \sum_{j\in\mathcal{V}} \rho_r(i) \left(\boldsymbol{D}^{-1}\boldsymbol{A}\right)_{ij}^{L} \beta_j. \tag{36}$$

□

A.2. Proof of Lemma 4.2

Proof. For $j \in S^{(b)}$, we can re-express $\widehat{\boldsymbol{P}}^L_{ij} = \left(\widehat{\boldsymbol{P}}^{(b)}\right)^L_{ij} = \left(\boldsymbol{e}^{(i)}\right)^\intercal \left(\widehat{\boldsymbol{P}}^{(b)}\right)^L \boldsymbol{e}^{(j)}$[4]. By the spectral properties of $\widehat{\boldsymbol{P}}^{(b)}$, $\left(\boldsymbol{e}^{(i)}\right)^\intercal \boldsymbol{v}_1^{(b)} = \sqrt{\frac{\widehat{\boldsymbol{D}}_{ii}}{\text{vol}\left(\mathcal{G}^{(b)}\right)}}$ (Lovász, 2001). Hence:

$$\widehat{\boldsymbol{P}}^L_{ij} = \sum_{k=1}^{|S^{(b)}|} \left(\lambda_k^{(b)}\right)^L \left(\boldsymbol{e}^{(i)}\right)^\intercal \boldsymbol{v}_k^{(b)} \left(\boldsymbol{v}_k^{(b)}\right)^\intercal \boldsymbol{e}^{(j)} \tag{37}$$

$$= \frac{\sqrt{\widehat{\boldsymbol{D}}_{ii}\widehat{\boldsymbol{D}}_{jj}}}{\text{vol}\left(\mathcal{G}^{(b)}\right)} + \sum_{k=2}^{|S^{(b)}|} \left(\lambda_k^{(b)}\right)^L \left(\boldsymbol{e}^{(i)}\right)^\intercal \boldsymbol{v}_k^{(b)} \left(\boldsymbol{v}_k^{(b)}\right)^\intercal \boldsymbol{e}^{(j)} \tag{38}$$

Then, by Cauchy-Schwarz:

$$\left|\widehat{\boldsymbol{P}}^L_{ij} - \frac{\sqrt{\widehat{\boldsymbol{D}}_{ii}\widehat{\boldsymbol{D}}_{jj}}}{\text{vol}\left(\mathcal{G}^{(b)}\right)}\right| \leq \left(\lambda^{(b)}\right)^L \sum_{k=1}^{|S^{(b)}|} \left|\left(\boldsymbol{e}^{(i)}\right)^\intercal \boldsymbol{v}_k^{(b)}\right| \left|\left(\boldsymbol{e}^{(j)}\right)^\intercal \boldsymbol{v}_k^{(b)}\right| \tag{39}$$

$$\leq \left(\lambda^{(b)}\right)^L \left(\sum_{k=1}^{|S^{(b)}|} \left|\left(\boldsymbol{e}^{(i)}\right)^\intercal \boldsymbol{v}_k^{(b)}\right|^2\right)^{\frac{1}{2}} \left(\sum_{k=1}^{|S^{(b)}|} \left|\left(\boldsymbol{e}^{(j)}\right)^\intercal \boldsymbol{v}_k^{(b)}\right|^2\right)^{\frac{1}{2}} \tag{40}$$

$$= \left(\lambda^{(b)}\right)^L \left(\left(\boldsymbol{e}^{(i)}\right)^\intercal \boldsymbol{V}^{(b)} \left(\boldsymbol{V}^{(b)}\right)^\intercal \boldsymbol{e}^{(i)}\right)^{\frac{1}{2}} \left(\left(\boldsymbol{e}^{(j)}\right)^\intercal \boldsymbol{V}^{(b)} \left(\boldsymbol{V}^{(b)}\right)^\intercal \boldsymbol{e}^{(j)}\right)^{\frac{1}{2}} \tag{41}$$

$$= \left(\lambda^{(b)}\right)^L \left\|\boldsymbol{e}^{(i)}\right\|_2 \left\|\boldsymbol{e}^{(j)}\right\|_2 \tag{42}$$

$$= \left(\lambda^{(b)}\right)^L \tag{43}$$

Let $\boldsymbol{P}^L = \left(\widehat{\boldsymbol{P}} + \Xi^{(0)}\right)^L = \widehat{\boldsymbol{P}}^L + \Xi^{(L)}$. Then, by the triangle inequality:

$$\left|\boldsymbol{P}^L_{ij} - \frac{\sqrt{\widehat{\boldsymbol{D}}_{ii}\widehat{\boldsymbol{D}}_{jj}}}{\text{vol}\left(\mathcal{G}^{(b)}\right)}\right| \leq \left(\lambda^{(b)}\right)^L + \left|\left(\boldsymbol{e}^{(i)}\right)^\intercal \Xi^{(L)} \boldsymbol{e}^{(j)}\right| \tag{44}$$

$$\leq \left(\lambda^{(b)}\right)^L + \left\|\Xi^{(L)}\right\|_{op} \tag{45}$$

$$\leq \left(\lambda^{(b)}\right)^L + \sum_{l=1}^{L} \binom{L}{l} \left\|\Xi^{(0)}\right\|_{op}^{l} \left\|\widehat{\boldsymbol{P}}\right\|_{op}^{L-l} \tag{46}$$

For $j \notin S^{(b)}$, $\widehat{\boldsymbol{P}}^L_{ij} = 0$. Then:

$$\left|\boldsymbol{P}^L_{ij} - 0\right| \leq \left|\left(\boldsymbol{e}^{(i)}\right)^\intercal \Xi^{(L)} \boldsymbol{e}^{(j)}\right| \tag{47}$$

$$\leq \sum_{l=1}^{L} \binom{L}{l} \left\|\Xi^{(0)}\right\|_{op}^{l} \left\|\widehat{\boldsymbol{P}}\right\|_{op}^{L-l} \tag{48}$$

□

[4]For simplicity, we abuse notation here: $\left(\widehat{\boldsymbol{P}}^{(b)}\right)^L_{ij}$ is not the entry at row i and column j, but rather the entry at the row corresponding to node i and column corresponding to node j. Similarly, $\boldsymbol{e}^{(i)}$ is the standard basis vector with a 1 at the entry corresponding to node i.

A.3. Proof of Theorem 4.3

Proof. For $u, v \in \mathcal{V}$, let $|\delta_{uv}| \leq \zeta_s$. Combining Lemmas 4.1 and 4.2, by our assumption that the computation graph paths to i, j are activated independently:

$$\mathbb{E}\left[f_{LP}\left(\boldsymbol{s}_i^{(L)}, \boldsymbol{s}_j^{(L)}\right)\right] = \mathbb{E}\left[\boldsymbol{s}_i^{(L)}\right]^\intercal \mathbb{E}\left[\boldsymbol{s}_j^{(L)}\right] \tag{49}$$

$$= \overline{\rho}_s^2(b)\left(\sum_{k\in S^{(b)}} \frac{\sqrt{\widehat{\boldsymbol{D}}_{ii}\widehat{\boldsymbol{D}}_{kk}}}{\mathrm{vol}\left(\mathcal{G}^{(b)}\right)}\alpha_k + \sum_{k\in\mathcal{V}}\delta_{ik}\alpha_k\right)^\intercal\left(\sum_{k\in S^{(b)}} \frac{\sqrt{\widehat{\boldsymbol{D}}_{jj}\widehat{\boldsymbol{D}}_{kk}}}{\mathrm{vol}\left(\mathcal{G}^{(b)}\right)}\alpha_k + \sum_{k\in\mathcal{V}}\delta_{jk}\alpha_k\right) \tag{50}$$

$$= \overline{\rho}_s^2(b)\sqrt{\widehat{\boldsymbol{D}}_{ii}\widehat{\boldsymbol{D}}_{jj}}\underbrace{\left\|\sum_{k\in S^{(b)}}\frac{\sqrt{\widehat{\boldsymbol{D}}_{kk}}}{\mathrm{vol}(\mathcal{G}^{(b)})}\alpha_k\right\|_2^2}_{\geq 0} \tag{51}$$

$$+ \overline{\rho}_s^2(b)\left(\sqrt{\widehat{\boldsymbol{D}}_{ii}}\sum_{k\in S^{(b)}}\frac{\sqrt{\widehat{\boldsymbol{D}}_{kk}}}{\mathrm{vol}(\mathcal{G}^{(b)})}\alpha_k\right)^\intercal\left(\sum_{k\in\mathcal{V}}\delta_{jk}\alpha_k\right) \tag{52}$$

$$+ \overline{\rho}_s^2(b)\left(\sum_{k\in\mathcal{V}}\delta_{ik}\alpha_k\right)^\intercal\left(\sqrt{\widehat{\boldsymbol{D}}_{jj}}\sum_{k\in S^{(b)}}\frac{\sqrt{\widehat{\boldsymbol{D}}_{kk}}}{\mathrm{vol}(\mathcal{G}^{(b)})}\alpha_k\right) \tag{53}$$

$$+ \overline{\rho}_s^2(b)\left(\sum_{k\in\mathcal{V}}\delta_{ik}\alpha_k\right)^\intercal\left(\sum_{k\in\mathcal{V}}\delta_{jk}\alpha_k\right) \tag{54}$$

Then, by Cauchy-Schwarz and the triangle inequality:

$$\left|\mathbb{E}\left[f_{LP}\left(\boldsymbol{s}_i^{(L)}, \boldsymbol{s}_j^{(L)}\right)\right] - \underbrace{\overline{\rho}_s^2(b)\left\|\sum_{k\in S^{(b)}}\frac{\sqrt{\widehat{\boldsymbol{D}}_{kk}}}{\mathrm{vol}(\mathcal{G}^{(b)})}\alpha_k\right\|_2^2\sqrt{\widehat{\boldsymbol{D}}_{ii}\widehat{\boldsymbol{D}}_{jj}}}_{\propto\sqrt{\widehat{\boldsymbol{D}}_{ii}\widehat{\boldsymbol{D}}_{jj}}}\right| \tag{55}$$

$$\leq \zeta_s\overline{\rho}_s^2(b)\left(\sqrt{\widehat{\boldsymbol{D}}_{ii}} + \sqrt{\widehat{\boldsymbol{D}}_{jj}}\right)\left\|\sum_{k\in S^{(b)}}\frac{\sqrt{\widehat{\boldsymbol{D}}_{kk}}}{\mathrm{vol}(\mathcal{G}^{(b)})}\alpha_k\right\|_2\left(\sum_{k\in\mathcal{V}}\|\alpha_k\|_2\right) + \zeta_s^2\overline{\rho}_s^2(b)\left(\sum_{k\in\mathcal{V}}\|\alpha_k\|_2\right)^2 \tag{56}$$

□

A.4. Lemma A.1 and Proof

Lemma A.1. *We introduce the notation $\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{A}$. We further define $\widehat{\boldsymbol{P}} = \widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{A}}$. Fix $i \in S^{(b)}$. Then, for $j \in S^{(b)}$:*

$$\left| \boldsymbol{P}_{ij}^{L} - \frac{\widehat{\boldsymbol{D}}_{jj}}{vol\left(\mathcal{G}^{(b)}\right)} \right| \leq \sqrt{\frac{\widehat{\boldsymbol{D}}_{jj}}{\widehat{\boldsymbol{D}}_{ii}}} \left(\lambda^{(b)}\right)^{L} + \sum_{l=1}^{L} \binom{L}{l} \left\| \Xi^{(0)} \right\|_{op}^{l} \left\| \widehat{\boldsymbol{P}} \right\|_{op}^{L-l} \tag{57}$$

And for $j \notin S^{(b)}$:

$$\left| \boldsymbol{P}_{ij}^{L} - 0 \right| \leq \sum_{l=1}^{L} \binom{L}{l} \left\| \Xi^{(0)} \right\|_{op}^{l} \left\| \widehat{\boldsymbol{P}} \right\|_{op}^{L-l} \tag{58}$$

Proof. Similar to the proof of Lemma 4.2:

$$\widehat{\boldsymbol{P}}_{ij}^{L} = \frac{\widehat{\boldsymbol{D}}_{jj}}{\mathrm{vol}\left(\mathcal{G}^{(b)}\right)} + \sqrt{\frac{\widehat{\boldsymbol{D}}_{jj}}{\widehat{\boldsymbol{D}}_{ii}}} \sum_{k=2}^{\left|S^{(b)}\right|} \left(\lambda_{k}^{(b)}\right)^{L} \left(\boldsymbol{e}^{(i)}\right)^{\intercal} \boldsymbol{v}_{k}^{(b)} \left(\boldsymbol{v}_{k}^{(b)}\right)^{\intercal} \boldsymbol{e}^{(j)} \tag{59}$$

Subsequently:

$$\left| \widehat{\boldsymbol{P}}_{ij}^{L} - \frac{\widehat{\boldsymbol{D}}_{jj}}{\mathrm{vol}\left(\mathcal{G}^{(b)}\right)} \right| \leq \sqrt{\frac{\widehat{\boldsymbol{D}}_{jj}}{\widehat{\boldsymbol{D}}_{ii}}} \left(\lambda^{(b)}\right)^{L} \tag{60}$$

Finally:

$$\left| \boldsymbol{P}_{ij}^{L} - \frac{\widehat{\boldsymbol{D}}_{jj}}{\mathrm{vol}\left(\mathcal{G}^{(b)}\right)} \right| \leq \zeta_r = \max_{u,v \in \mathcal{V}} \sqrt{\frac{\widehat{\boldsymbol{D}}_{vv}}{\widehat{\boldsymbol{D}}_{uu}}} \left(\lambda^{(b)}\right)^{L} + \sum_{l=1}^{L} \binom{L}{l} \left\| \Xi^{(0)} \right\|_{op}^{l} \left\| \widehat{\boldsymbol{P}} \right\|_{op}^{L-l} \tag{61}$$

For $j \notin S^{(b)}$, $\widehat{\boldsymbol{P}}_{ij}^{L} = 0$. Then:

$$\left| \boldsymbol{P}_{ij}^{L} - 0 \right| \leq \sum_{l=1}^{L} \binom{L}{l} \left\| \Xi^{(0)} \right\|_{op}^{l} \left\| \widehat{\boldsymbol{P}} \right\|_{op}^{L-l} \leq \zeta_r \tag{62}$$

□

A.5. Proof of Theorem 4.4

Proof. For $u, v \in \mathcal{V}$, let $|\delta_{uv}| \leq \zeta_r$. Combining Lemmas 4.1 and A.1, by our assumption that the computation graph paths to i, j are activated independently:

$$\mathbb{E}\left[f_{LP}\left(\boldsymbol{r}_i^{(L)}, \boldsymbol{r}_j^{(L)}\right)\right] = \mathbb{E}\left[\boldsymbol{r}_i^{(L)}\right]^\intercal \mathbb{E}\left[\boldsymbol{r}_j^{(L)}\right] \tag{63}$$

$$= \overline{\rho}_r^2(b)\left(\sum_{k\in S^{(b)}} \frac{\widehat{\boldsymbol{D}}_{kk}}{\text{vol}\left(\mathcal{G}^{(b)}\right)}\beta_k + \sum_{k\in\mathcal{V}}\delta_{ik}\beta_k\right)^\intercal \left(\sum_{k\in S^{(b)}} \frac{\widehat{\boldsymbol{D}}_{kk}}{\text{vol}\left(\mathcal{G}^{(b)}\right)}\beta_k + \sum_{k\in\mathcal{V}}\delta_{jk}\beta_k\right) \tag{64}$$

$$= \overline{\rho}_r^2(b)\underbrace{\left\|\sum_{k\in S^{(b)}} \frac{\widehat{\boldsymbol{D}}_{kk}}{\text{vol}(\mathcal{G}^{(b)})}\beta_k\right\|_2^2}_{\geq 0} + \overline{\rho}_r^2(b)\left(\sum_{k\in S^{(b)}} \frac{\widehat{\boldsymbol{D}}_{kk}}{\text{vol}(\mathcal{G}^{(b)})}\beta_k\right)^\intercal \left(\sum_{k\in\mathcal{V}}\delta_{jk}\beta_k\right) \tag{65}$$

$$+ \overline{\rho}_r^2(b)\left(\sum_{k\in\mathcal{V}}\delta_{ik}\beta_k\right)^\intercal \left(\sum_{k\in S^{(b)}} \frac{\widehat{\boldsymbol{D}}_{kk}}{\text{vol}(\mathcal{G}^{(b)})}\beta_k\right) + \overline{\rho}_r^2(b)\left(\sum_{k\in\mathcal{V}}\delta_{ik}\beta_k\right)^\intercal \left(\sum_{k\in\mathcal{V}}\delta_{jk}\beta_k\right) \tag{66}$$

Then, by Cauchy-Schwarz and the triangle inequality:

$$\left|\mathbb{E}\left[f_{LP}\left(\boldsymbol{r}_i^{(L)}, \boldsymbol{r}_j^{(L)}\right)\right] - \underbrace{\overline{\rho}_r^2(b)\left\|\sum_{k\in S^{(b)}} \frac{\widehat{\boldsymbol{D}}_{kk}}{\text{vol}(\mathcal{G}^{(b)})}\beta_k\right\|_2^2}_{\propto \text{ constant}}\right| \tag{67}$$

$$\leq \zeta_r\overline{\rho}_r^2(b)\left\|\sum_{k\in S^{(b)}} \frac{\widehat{\boldsymbol{D}}_{kk}}{\text{vol}(\mathcal{G}^{(b)})}\beta_k\right\|_2\left(\sum_{k\in\mathcal{V}}\|\beta_k\|_2\right) + \zeta_r^2\overline{\rho}_r^2(b)\left(\sum_{k\in\mathcal{V}}\|\beta_k\|_2\right)^2 \tag{68}$$

□

B. Approximation of $\Delta^{(b)}$

B.1. Approximation of $\Delta^{(b)}$ for Φ_s

$$\Delta^{(b)}\left(\boldsymbol{s}_i^{(L)}, \boldsymbol{s}_j^{(L)}\right) \tag{69}$$

$$= \Big| \frac{1}{|(S^{(b)} \cap T^{(1)}) \times S^{(b)}|} \sum_{i \in S^{(b)} \cap T^{(1)}} \sum_{j \in S^{(b)}} f_{LP}\left(\boldsymbol{s}_i^{(L)}, \boldsymbol{s}_j^{(L)}\right) \tag{70}$$

$$- \frac{1}{|(S^{(b)} \cap T^{(2)}) \times S^{(b)}|} \sum_{i \in S^{(b)} \cap T^{(2)}} \sum_{j \in S^{(b)}} f_{LP}\left(\boldsymbol{s}_i^{(L)}, \boldsymbol{s}_j^{(L)}\right) \Big| \tag{71}$$

$$\approxeq \Big| \frac{1}{|S^{(b)} \cap T^{(1)}|\,|S^{(b)}|} \sum_{i \in S^{(b)} \cap T^{(1)}} \sum_{j \in S^{(b)}} \overline{\rho}_s^2(b) \sqrt{\widehat{\boldsymbol{D}}_{ii} \widehat{\boldsymbol{D}}_{jj}} \left\| \sum_{k \in S^{(b)}} \frac{\sqrt{\widehat{\boldsymbol{D}}_{kk}}}{\mathrm{vol}(\mathcal{G}^{(b)})} \alpha_k \right\|_2^2 \tag{72}$$

$$- \frac{1}{|S^{(b)} \cap T^{(2)}|\,|S^{(b)}|} \sum_{i \in S^{(b)} \cap T^{(2)}} \sum_{j \in S^{(b)}} \overline{\rho}_s^2(b) \sqrt{\widehat{\boldsymbol{D}}_{ii} \widehat{\boldsymbol{D}}_{jj}} \left\| \sum_{k \in S^{(b)}} \frac{\sqrt{\widehat{\boldsymbol{D}}_{kk}}}{\mathrm{vol}(\mathcal{G}^{(b)})} \alpha_k \right\|_2^2 \Big| \tag{73}$$

$$= \frac{\overline{\rho}_s^2(b)}{|S^{(b)}|} \left\| \sum_{k \in S^{(b)}} \frac{\sqrt{\widehat{\boldsymbol{D}}_{kk}}}{\mathrm{vol}(\mathcal{G}^{(b)})} \alpha_k \right\|_2^2 \left| \sum_{j \in S^{(b)}} \sqrt{\widehat{\boldsymbol{D}}_{jj}} \underbrace{\left(\mathop{\mathbb{E}}_{i \sim U(S^{(b)} \cap T^{(1)})} \sqrt{\widehat{\boldsymbol{D}}_{ii}} - \mathop{\mathbb{E}}_{i \sim U(S^{(b)} \cap T^{(2)})} \sqrt{\widehat{\boldsymbol{D}}_{ii}} \right)}_{\text{degree disparity}} \right| \tag{74}$$

B.2. Approximation of $\Delta^{(b)}$ for Φ_r

$$\Delta^{(b)}\left(\boldsymbol{r}_i^{(L)}, \boldsymbol{r}_j^{(L)}\right) \tag{75}$$

$$= \Big| \frac{1}{|(S^{(b)} \cap T^{(1)}) \times S^{(b)}|} \sum_{i \in S^{(b)} \cap T^{(1)}} \sum_{j \in S^{(b)}} f_{LP}\left(\boldsymbol{r}_i^{(L)}, \boldsymbol{r}_j^{(L)}\right) \tag{76}$$

$$- \frac{1}{|(S^{(b)} \cap T^{(2)}) \times S^{(b)}|} \sum_{i \in S^{(b)} \cap T^{(2)}} \sum_{j \in S^{(b)}} f_{LP}\left(\boldsymbol{r}_i^{(L)}, \boldsymbol{r}_j^{(L)}\right) \Big| \tag{77}$$

$$\approxeq \Big| \frac{1}{|S^{(b)} \cap T^{(1)}|\,|S^{(b)}|} \sum_{i \in S^{(b)} \cap T^{(1)}} \sum_{j \in S^{(b)}} \overline{\rho}_r^2(b) \left\| \sum_{k \in S^{(b)}} \frac{\widehat{\boldsymbol{D}}_{kk}}{\mathrm{vol}(\mathcal{G}^{(b)})} \beta_k \right\|_2^2 \tag{78}$$

$$- \frac{1}{|S^{(b)} \cap T^{(2)}|\,|S^{(b)}|} \sum_{i \in S^{(b)} \cap T^{(2)}} \sum_{j \in S^{(b)}} \overline{\rho}_r^2(b) \left\| \sum_{k \in S^{(b)}} \frac{\widehat{\boldsymbol{D}}_{kk}}{\mathrm{vol}(\mathcal{G}^{(b)})} \beta_k \right\|_2^2 \Big| \tag{79}$$

$$= 0 \tag{80}$$

C. Datasets Used in §6.1

In our experiments in §6.1, we use 10 real-world network datasets from Bojchevski & Günnemann (2018), Shchur et al. (2018), Rozemberczki & Sarkar (2020), and Rozemberczki et al. (2021), covering diverse domains (e.g., citation networks, collaboration networks, online social networks). We provide a description and some statistics of each dataset in Table 2. All the datasets have node features and are undirected. We were unable to find the exact class names and their label correspondence from the dataset documentation.

- In all the citation network datasets, nodes represent documents, edges represent citation links, and features are a bag-of-words representation of documents. We row-normalize the features to sum to 1, following Fey & Lenssen (2019)[5]. The classification task is to predict the topic of documents.
- In the collaboration network datasets, nodes represent authors, edges represent coauthorships, and features are embeddings of paper keywords for authors' papers. The classification task is to predict the most active field of study for authors.
- In the LastFMAsia network dataset, nodes represent LastFM users from Asia, edges represent friendships between users, and features are embeddings of the artists liked by users. The classification task is to predict the home country of users.
- In the Twitch network datasets, nodes represent gamers on Twitch, edges represent followerships between them, and features are embeddings of the history of games played by the Twitch users. The classification task is to predict whether or not a gamer streams adult content.

We only run experiments on datasets that can fit without sampling nodes on a single NVIDIA GeForce GTX Titan Xp Graphic Card with 12196MiB of space. Furthermore, we only consider the three largest datasets (i.e., with the most nodes) from Rozemberczki et al. (2021). We use PyTorch Geometric to load and process all datasets (Fey & Lenssen, 2019).

Table 2: Summary of the datasets used in our experiments.

Name	Domain	# Nodes	# Edges	# Features	# Classes
Cora	citation	19793	126842	8710	70
CiteSeer	citation	4230	10674	602	6
DBLP	citation	17716	105734	1639	4
PubMed	citation	19717	88648	500	3
CS	collaboration	18333	163788	6805	15
Physics	collaboration	34493	495924	8415	5
LastFMAsia	online social	7624	55612	128	18
Twitch-DE	online social	9498	315774	128	2
Twitch-EN	online social	7126	77774	128	2
Twitch-FR	online social	6551	231883	128	2

[5] https://github.com/pyg-team/pytorch_geometric/blob/master/examples/link_pred.py

D. Datasets Used in §6.2

We run experiments on three network datasets: (1) the NBA social network (cf. §D.1), (2) the German credit network (cf. §D.2), and (3) a new DBLP-Fairness citation network that we construct (cf. §D.3). All the datasets have node features and are undirected. We do not pass sensitive attributes as features to the models that we train. For each dataset, we min-max normalize node features to fall in $[-1, 1]$, following Dai & Wang (2021) and Agarwal et al. (2021). Furthermore, for all datasets, $D = 2$.

D.1. NBA Dataset

The NBA network (Dai & Wang, 2021) has 403 nodes representing NBA basketball players who are connected if they follow each other on Twitter. There are 21242 links. Each node has 95 features, with an average degree of 52.71 ± 35.14. We consider two sensitive attributes per node:

- Age $\{S^{(b)}\}_{b\in[B]}$: how old the payer is, i.e., YOUNG (≤ 25 years) or OLD (> 25 years).
- Nationality $\{T^{(d)}\}_{d\in[D]}$: from where the player is, i.e., UNITED STATES or OVERSEAS.

D.2. German Dataset

The German network (Agarwal et al., 2021) comprises 1000 nodes representing clients in a German bank who are connected if they have similar credit accounts. The German network is not natively a graph dataset; synthetic edges were created by Agarwal et al. There are 44484 links. Each node has 27 features (e.g., loan amount, account-related features), with an average degree of 44.48 ± 26.52. We consider two sensitive attributes per node:

- Foreign worker $\{S^{(b)}\}_{b\in[B]}$: whether the client is a foreign worker, i.e., YES or NO.
- Gender $\{T^{(d)}\}_{d\in[D]}$: the gender of the client, i.e., MAN or WOMAN.

D.3. DBLP-Fairness Dataset

In this subsection, we detail how we construct the DBLP-Fairness dataset. We build DBLP-Fairness, as there are only a few natively-graph network datasets with sensitive attributes that are appropriate for graph learning (Subramonian et al., 2022).

We begin with the version of the DBLP-Citation-network V12 dataset from (Tang et al., 2008) that was processed by Xu et al. (2021). This dataset has 3658127 nodes. Each node represents a paper and each edge represents a citation link. We consider five node features:

- Team size: the number of authors on the paper.
- Mean collaborators: the average number of collaborators with whom the authors have previously published.
- Gini collaborators: the Gini coefficient of the number of collaborators with whom the authors have previously published.
- Mean productivity: the average number of papers that the authors have previously published.
- Gini productivity: the Gini coefficient of the number of papers that the authors have previously published.

We also consider two sensitive attributes per node:

- Field $\{S^{(b)}\}_{b\in[B]}$: the field to which the paper belongs, i.e., PROGRAMMING LANGUAGES or DATABASES.
- Nationality $\{T^{(d)}\}_{d\in[D]}$: the country where most authors reside, i.e., UNITED STATES or CHINA.

In DBLP-Fairness, we only include papers whose nationality is UNITED STATES or CHINA; American and Chinese citation networks are known to be stratified (Zhao et al., 2022). We also only include papers whose field is PROGRAMMING LANGUAGES or DATABASES; we infer the field of a paper using its keywords (i.e., whether they contain "programming language" and "database"), and discard papers which include both "programming language" and "database" in its keywords. Furthermore, we filter out all papers from before 2010. We sought DBLB-Fairness to be of comparable size to the citation networks in §C. Following filtering, we were left with 14537 nodes and 24844 edges.

E. Models

For all experiments, we use GCN encoders (Kipf & Welling, 2017) to get node representations. Each encoder has two layers (128-dimensional hidden layer, 64-dimensional output layer) with a ReLU nonlinearity in between. We only use two layers, as this is common practice in graph deep learning to prevent oversmoothing (Oono & Suzuki, 2020); however, we run experiments with four layers in §G. We do not use any regularization (e.g., Dropout, BatchNorm). The encoders are explicitly trained for LP with the inner-product LP score function in Eqn. 6, binary cross-entropy loss, and the Adam optimizer with full-batch gradient descent and a learning rate of 0.01 (Kingma & Ba, 2014). We use a random link split of 0.85-0.05-0.1 for train-val-test, following the PyTorch Geometric LP example[6]. We train the encoders for 100 epochs, with a new round of negative link sampling during every epoch; we use a 1:1 ratio of positive to negative links. We ultimately select the model parameters with the highest validation ROC-AUC. Although we do not do any hyperparameter tuning, the test ROC-AUC values (displayed in the figures in §6) indicate that the encoders are well-trained. We use PyTorch (Paszke et al., 2019) and PyTorch Geometric (Fey & Lenssen, 2019) to train all the encoders on a single NVIDIA GeForce GTX Titan Xp Graphic Card with 12196MiB of space.

[6] https://github.com/pyg-team/pytorch_geometric/blob/master/examples/link_pred.py

F. Remaining Plots

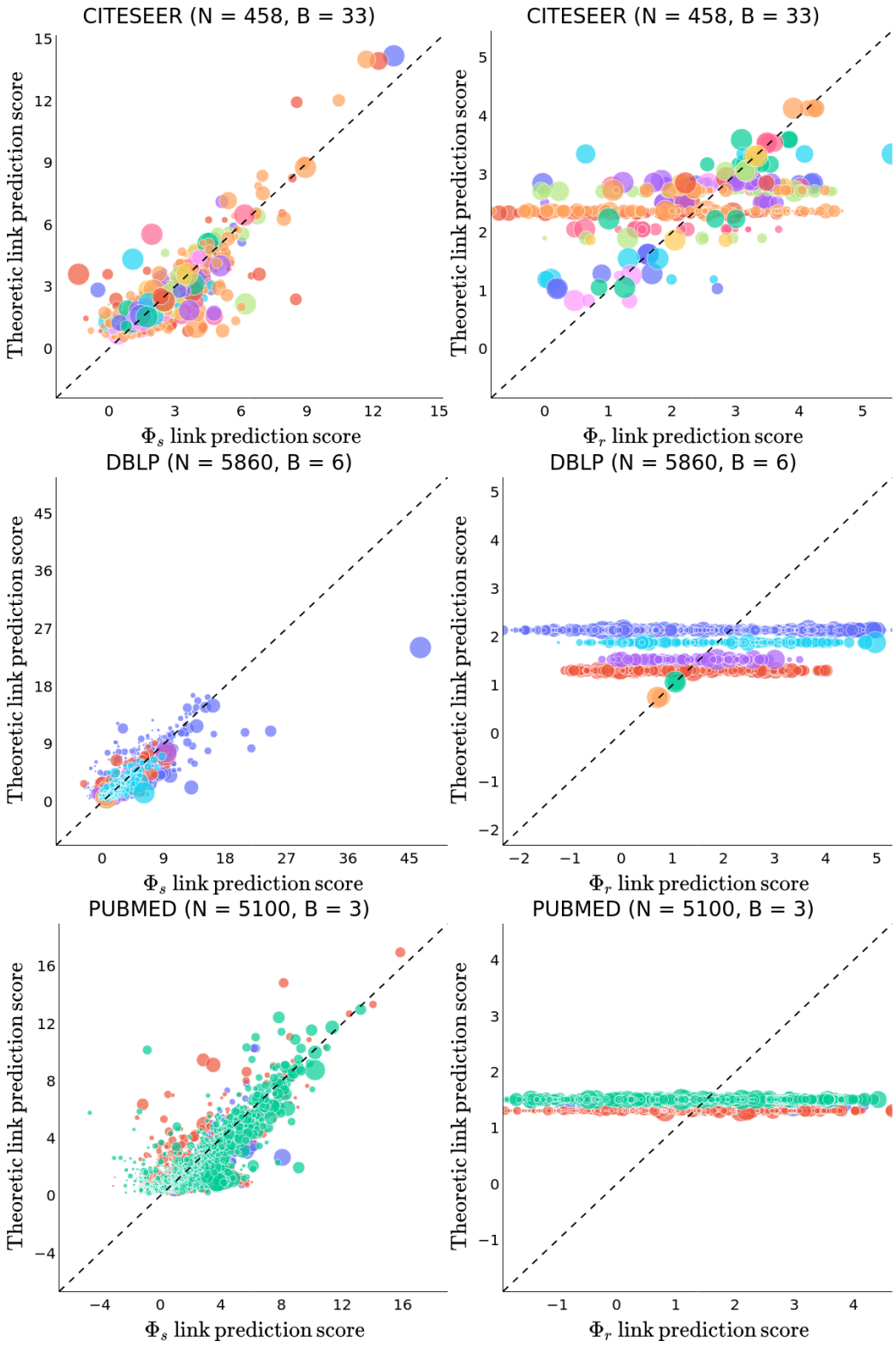


Figure 4: Theoretic vs. GCN LP scores for citation network datasets.

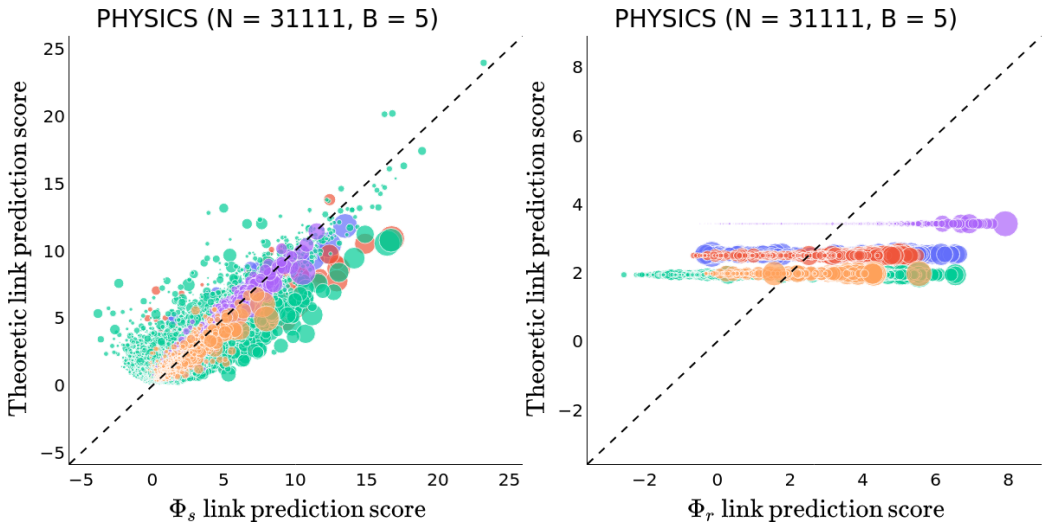


Figure 5: Theoretic vs. GCN LP scores for collaboration network datasets.

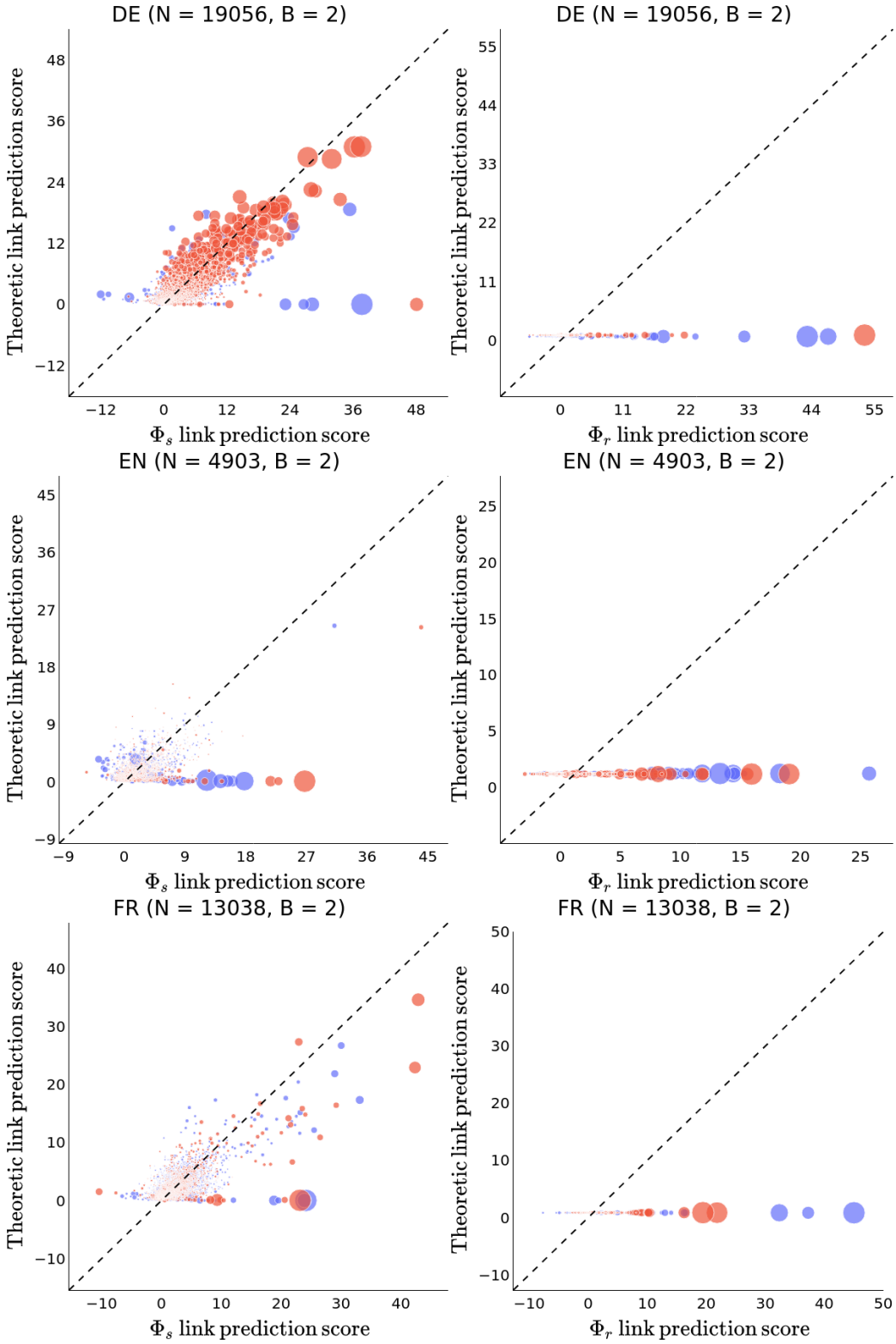


Figure 6: Theoretic vs. GCN LP scores for online social network datasets.

G. Additional Experiments

G.1. Additional Experiments for §6.1 (4-layer Encoders)

We run the experiments from §6.1 for Φ_s with the same settings, except we use 4-layer (instead of 2-layer) encoders (128-dimensional hidden layers, 64-dimensional output layer). We run these additional experiments because the error bound for the theoretic LP scores for Φ_s depends on the number of encoder layers L. We find that the experimental results continue to support our theoretical analysis, both qualitatively and quantitatively (cf. Table 3, Figure 7); the NRMSE and PCC values are comparable to or better than those from the experiments with the 2-layer encoders (especially for the EN dataset).

Table 3: The test AUC of the 4-layer Φ_s encoders on the real-world network datasets, and the NRMSE and PCC of the theoretic LP scores as predictors of the Φ_s scores.

	NRMSE (↓)	**PCC** (↑)	Φ_s **Test AUC** (↑)
CORA	0.044 ± 0.006	0.858 ± 0.026	0.853 ± 0.028
CITESEER	0.057 ± 0.006	0.890 ± 0.017	0.861 ± 0.026
DBLP	0.021 ± 0.002	0.885 ± 0.054	0.887 ± 0.019
PUBMED	0.056 ± 0.009	0.802 ± 0.024	0.900 ± 0.006
CS	0.039 ± 0.006	0.918 ± 0.008	0.949 ± 0.004
PHYSICS	0.030 ± 0.002	0.077 ± 0.013	0.950 ± 0.004
LASTFMASIA	0.040 ± 0.004	0.938 ± 0.005	0.949 ± 0.002
DE	0.014 ± 0.003	0.918 ± 0.025	0.882 ± 0.002
EN	0.034 ± 0.005	0.752 ± 0.036	0.846 ± 0.008
FR	0.019 ± 0.003	0.833 ± 0.038	0.896 ± 0.003

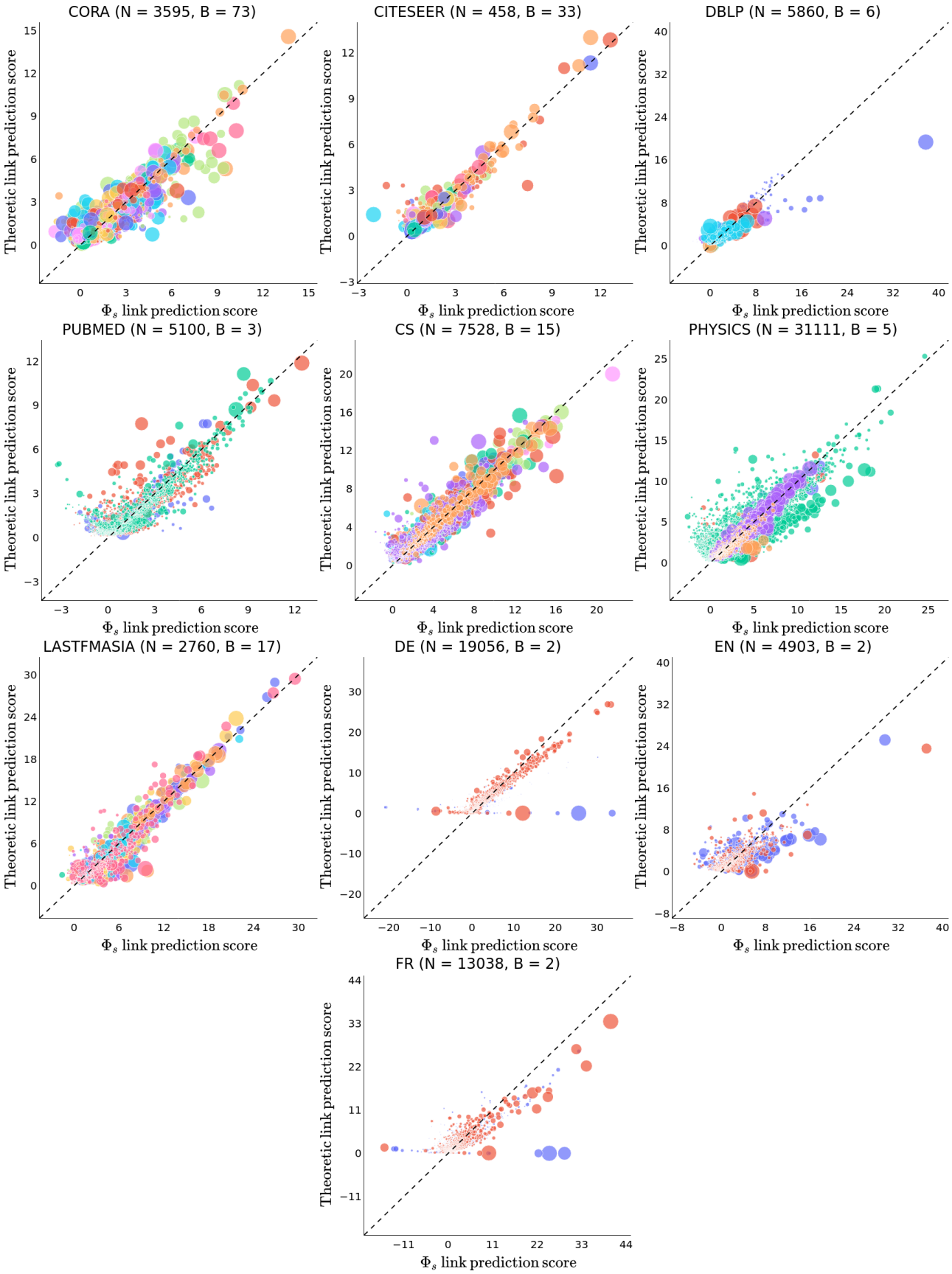


Figure 7: Theoretic LP score vs. 4-layer Φ_s LP score for all network datasets.

G.2. Additional Experiments for §6.1 (Hadamard Product and MLP LP Score Function)

We also run the experiments from §6.1 for Φ_s with the same settings, except we use the following LP score function:

$$f_{LP}\left(\boldsymbol{h}_i^{(L)}, \boldsymbol{h}_j^{(L)}\right) = f_{MLP}\left(\boldsymbol{h}_i^{(L)} \odot \boldsymbol{h}_j^{(L)}\right), \tag{81}$$

where $\odot$ is the Hadamard product and f_{MLP} is a 2-layer MLP with a 64-dimensional hidden layer and ReLU nonlinearity. We run these additional experiments because a Hadamard product and MLP score function is often used in the literature. We find that that our theoretical analysis is still relevant to and reasonably supports the experimental results, both qualitatively and quantitatively (cf. Table 4, Figure 8). This could be because MLPs have an inductive bias towards learning simpler, often linear functions (Nakkiran et al., 2019; Valle-Pérez et al., 2019), and our theoretical findings are generalizable to linear LP score functions. Notably, in this setting, Φ_s makes a higher number of negative link predictions. For a few datasets (e.g., Cora, CiteSeer, LastFMAsia), a handful of theoretic LP scores are negative because the regression (incorrectly) predicts $\overline{\rho}_s^2(b)$ for 1-2 groups $S^{(b)}$ to be negative.

Table 4: The test AUC of the Φ_s encoders with an f_{MLP} score function on the real-world network datasets, and the NRMSE and PCC of the theoretic LP scores as predictors of the Φ_s scores.

	NRMSE (↓)	**PCC** (↑)	Φ_s **Test AUC** (↑)
CORA	0.034 ± 0.004	0.830 ± 0.015	0.915 ± 0.001
CITESEER	0.090 ± 0.014	0.365 ± 0.070	0.913 ± 0.008
DBLP	0.026 ± 0.003	0.652 ± 0.029	0.933 ± 0.004
PUBMED	0.054 ± 0.007	0.813 ± 0.038	0.932 ± 0.003
CS	0.047 ± 0.008	0.677 ± 0.036	0.970 ± 0.001
PHYSICS	0.055 ± 0.007	0.566 ± 0.026	0.976 ± 0.001
LASTFMASIA	0.049 ± 0.008	0.682 ± 0.035	0.960 ± 0.003
DE	0.030 ± 0.008	0.683 ± 0.047	0.935 ± 0.001
EN	0.039 ± 0.006	0.463 ± 0.022	0.905 ± 0.002
FR	0.031 ± 0.006	0.654 ± 0.067	0.935 ± 0.002

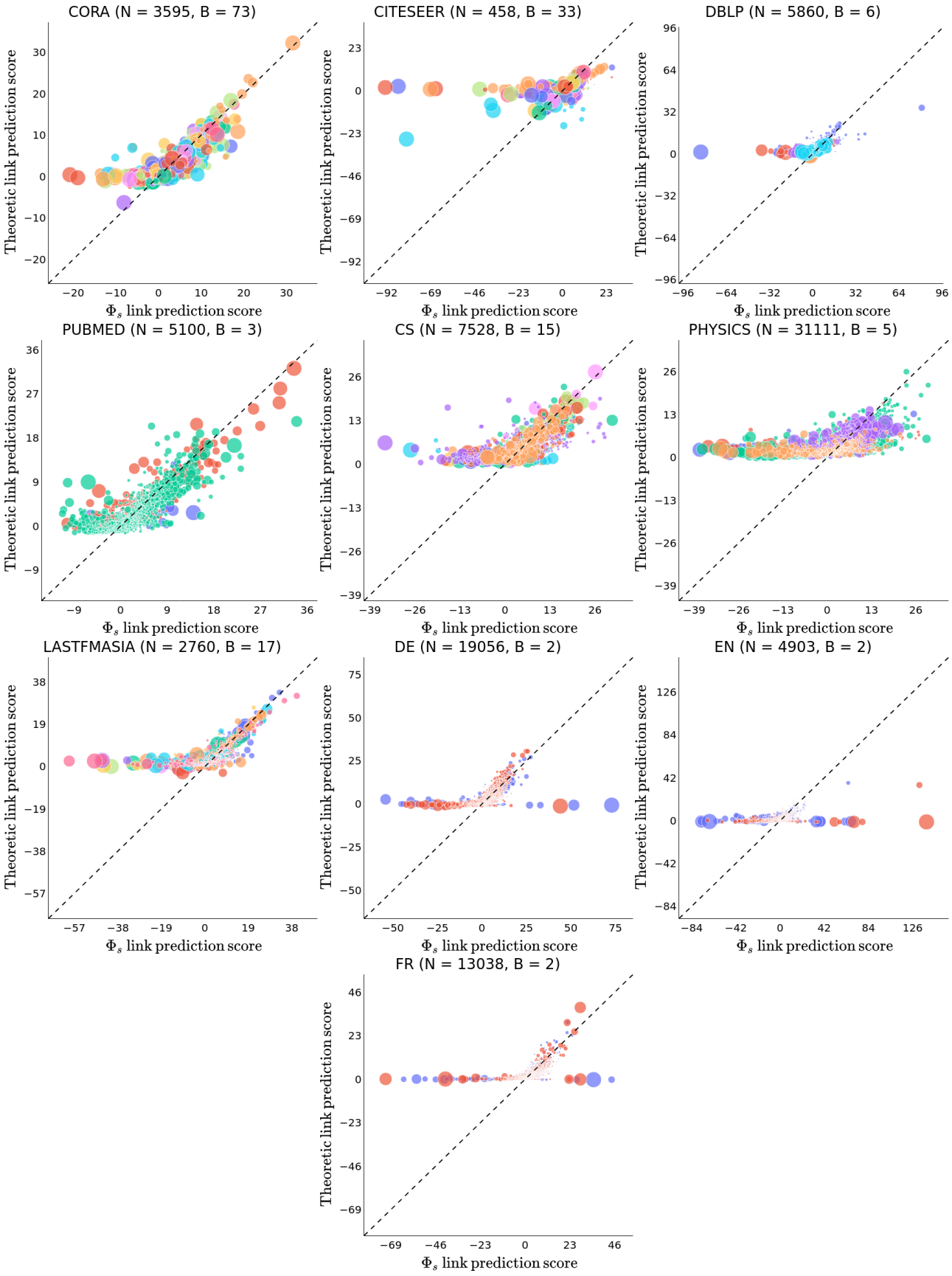


Figure 8: Theoretic LP score vs. Φ_s LP score (with Hadamard product and MLP) for all network datasets.

G.3. Additional Experiments for §6.2

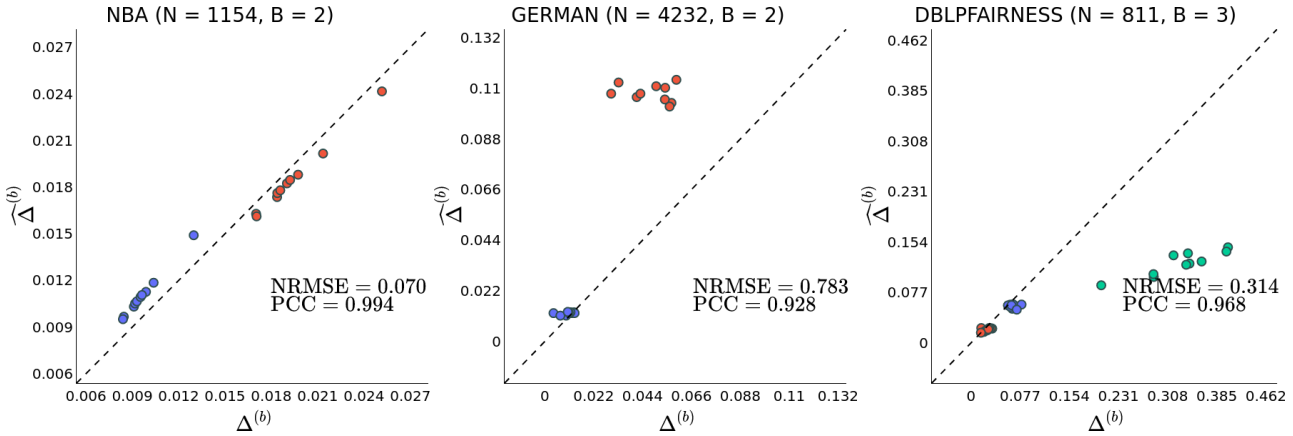


Figure 9: The plots display $\widehat{\Delta}^{(b)}$ vs. $\Delta^{(b)}$ for 4-layer Φ_s for the NBA, German, and DBLP-Fairness datasets over all $b \in [B]$ and 10 random seeds.

G.4. Additional Experiments for §6.3

Table 5: $\frac{1}{B}\sum_{b\in[B]} \Delta^{(b)}$ and the test AUC for the NBA, German, and DBLP-Fairness datasets with various settings of λ_{fair}. The **left** table corresponds to 4-layer Φ_s, and the **right** to 4-layer Φ_r.

	λ_{fair}	$\frac{1}{B}\sum_{b\in[B]} \Delta^{(b)}$ (↓)	Φ_s **Test AUC** (↑)
NBA	4.0	0.000 ± 0.000	0.752 ± 0.001
NBA	2.0	0.006 ± 0.001	0.752 ± 0.001
NBA	1.0	0.011 ± 0.001	0.753 ± 0.001
NBA	0.0	0.014 ± 0.001	0.753 ± 0.001
DBLPFAIRNESS	4.0	0.090 ± 0.041	0.793 ± 0.009
DBLPFAIRNESS	2.0	0.070 ± 0.015	0.800 ± 0.007
DBLPFAIRNESS	1.0	0.099 ± 0.009	0.804 ± 0.007
DBLPFAIRNESS	0.0	0.122 ± 0.028	0.820 ± 0.009
GERMAN	4.0	0.012 ± 0.008	0.817 ± 0.004
GERMAN	2.0	0.018 ± 0.007	0.827 ± 0.015
GERMAN	1.0	0.018 ± 0.008	0.856 ± 0.025
GERMAN	0.0	0.028 ± 0.007	0.874 ± 0.011

	λ_{fair}	$\frac{1}{B}\sum_{b\in[B]} \Delta^{(b)}$ (↓)	Φ_r **Test AUC** (↑)
NBA	4.0	0.000 ± 0.000	0.581 ± 0.029
NBA	2.0	0.000 ± 0.000	0.574 ± 0.021
NBA	1.0	0.000 ± 0.000	0.580 ± 0.025
NBA	0.0	0.000 ± 0.000	0.589 ± 0.031
DBLPFAIRNESS	4.0	0.034 ± 0.012	0.769 ± 0.009
DBLPFAIRNESS	2.0	0.045 ± 0.021	0.788 ± 0.007
DBLPFAIRNESS	1.0	0.074 ± 0.013	0.797 ± 0.006
DBLPFAIRNESS	0.0	0.095 ± 0.015	0.811 ± 0.006
GERMAN	4.0	0.027 ± 0.009	0.765 ± 0.013
GERMAN	2.0	0.023 ± 0.007	0.765 ± 0.011
GERMAN	1.0	0.031 ± 0.010	0.786 ± 0.030
GERMAN	0.0	0.030 ± 0.009	0.838 ± 0.025

H. Theory Pitfalls

To understand the second pitfall from §6.1, we separately investigate the association between the within-group degree product $\left(\widehat{\boldsymbol{D}}_{ii}\widehat{\boldsymbol{D}}_{jj}\right)$ and the absolute deviation of the theoretic LP scores from the Φ_s scores, as well as the association between the (transformed) feature similarity $\left(\left\|\sum_{k\in S^{(b)}} \frac{\sqrt{\widehat{\boldsymbol{D}}_{kk}}}{\mathrm{vol}(\mathcal{G}^{(b)})}\alpha_k\right\|_2^2\right)$ and the absolute deviation (cf. Figure 10). We observe that the absolute deviation is highest for the node pairs with a relatively small degree product (i.e., nodes with a low PA score) and low feature similarity.

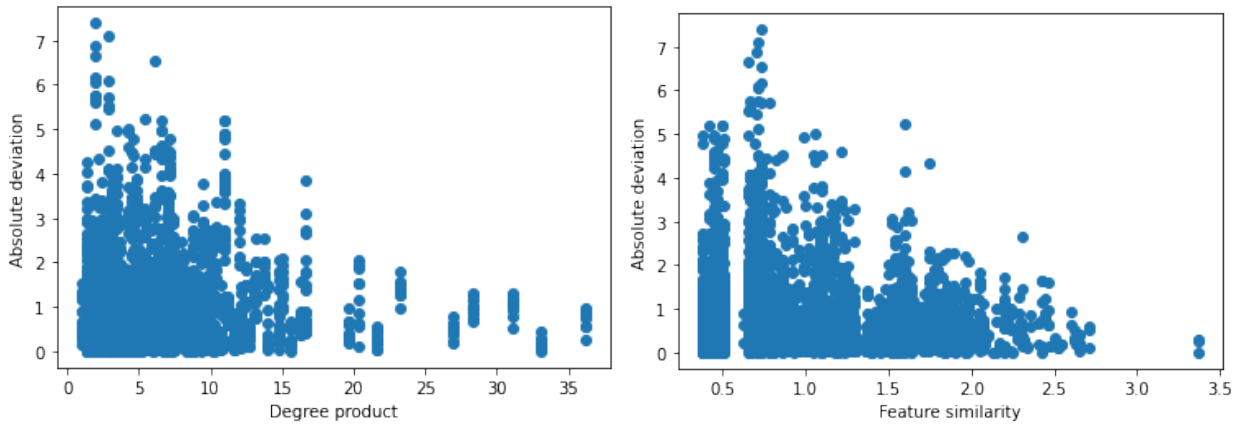


Figure 10: Associations of absolute deviation with degree product and with feature similarity for CiteSeer.

I. Error Analysis of Φ_r Theoretic Scores

Figure 11 reveals that the max term $\max_{u,v \in \mathcal{V}} \sqrt{\frac{\widehat{\boldsymbol{D}}_{vv}}{\widehat{\boldsymbol{D}}_{uu}}}$ is quite large in practice, which causes the theoretic LP scores to generally be poor estimates for the Φ_r scores. We additionally find in Figure 11 that the relative error (as measured by NRMSE and PCC) of the theoretic LP scores for Φ_r is not lower for lower values of the max term $\max_{u,v \in \mathcal{V}} \sqrt{\frac{\widehat{\boldsymbol{D}}_{vv}}{\widehat{\boldsymbol{D}}_{uu}}}$.

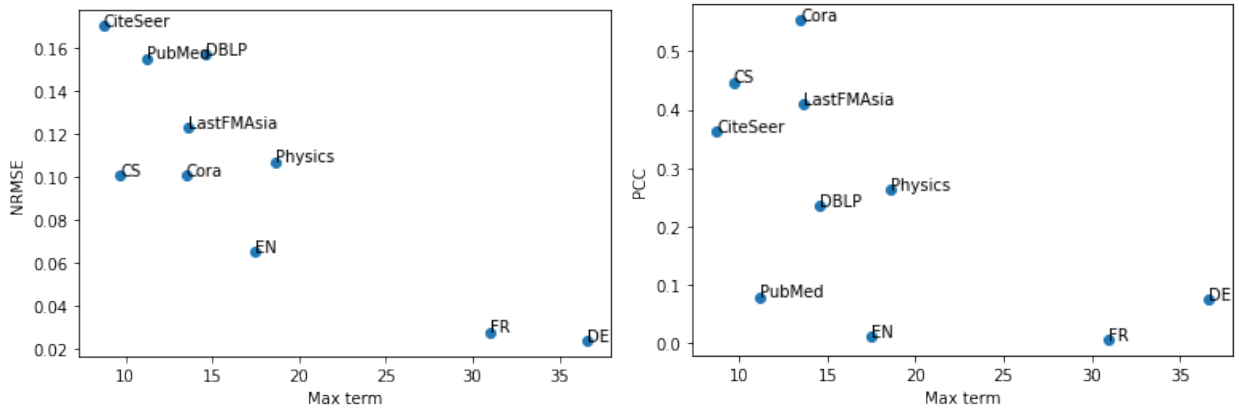


Figure 11: Weak associations of max term with NRMSE and PCC of theoretic LP scores for Φ_r across all datasets described in §C.

Furthermore, Figure 12 reveals that Φ_r LP scores are *not* higher for incident nodes with larger degrees.

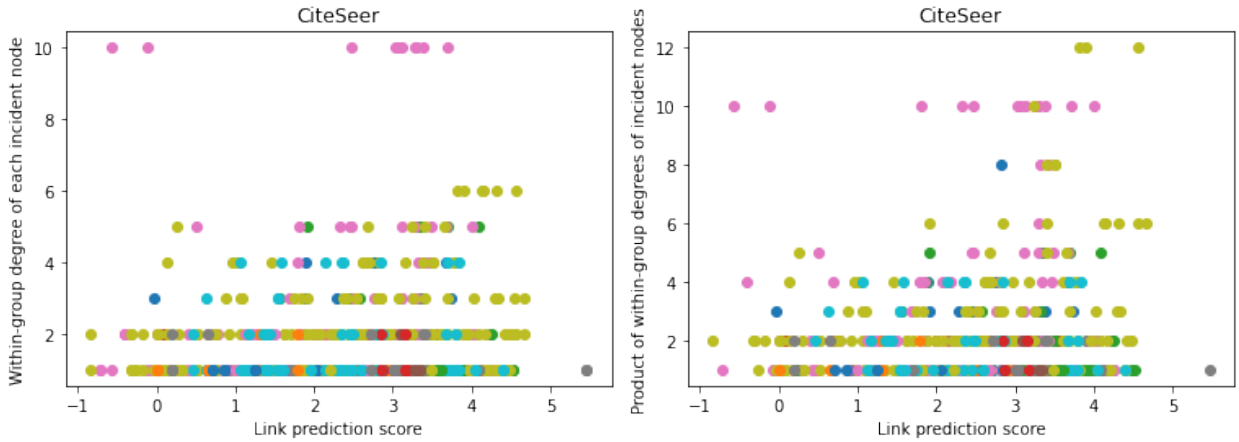


Figure 12: Weak associations of mean Φ_r LP scores (over 10 random seeds) with degree of each incident node and product of degrees of both incident nodes. Colors correspond to different groups.

There are intimate connections between Theorem 4.4 and the steady-state probabilities of random walks. The stationary probabilities of random walks are the same regardless of the starting node. This is why Φ_r produces similar representations for all the nodes in each social group, regardless of the degree of the node; in fact, with a larger number of layers, Φ_r would oversmooth all the representations to the same vector (Keriven, 2022). Hence, Φ_r LP scores do not have a degree dependence, theoretically or empirically.

J. Preferential Attachment and Motivation

Preferential Attachment Preferential attachment (PA) describes the propensity of links to form with high-degree nodes[7]. Network scientists have studied for decades how links in real-world networks exhibit PA. For example, in the iterative Barabási-Albert model of network formation, each new node s forms links with existing nodes t with probability proportional to the degree of t, i.e., $\mathbb{P}((s,t) \in \mathcal{E}) \propto deg(t)$. In the context of our paper, PA describes how a GCN with an inner-product LP score function often predicts links between nodes i, j with score $\propto \sqrt{deg(i) \cdot deg(j)}$ approximately (Theorem 4.3).

Motivation A wealth of literature in network science and the social sciences has examined the PA properties of real-world networks and how these properties contribute to unfair (non-neural) algorithms (§2). For example, Stoica et al. (2018) find that Instagram accounts run by men have a significantly higher following than those run by women due to gender discrimination; this degree disparity is only amplified by link recommendation algorithms that suggest following high-degree accounts, which makes the rich get richer and reveals that these algorithms have a PA bias. Moreover, many papers outside graph learning have discussed the intersectional unfairness of machine learning (§2).

However, despite the increasing real-world deployment of GNNs for LP, their unfairness has not been studied from the perspectives of PA and intersections of social groups. Our paper fills this gap by providing thorough theoretical and empirical evidence that GCNs (Kipf & Welling, 2017) have a PA bias when predicting links between nodes in the same social group. **This finding is nontrivial as GCNs leverage a combination of features and local structural context to make link predictions.**

Our research question is challenging from a technical perspective, as it requires uncovering properties of *short* random walks on graphs (since most GNNs are shallow); in contrast, most random walk results in the literature concern random walks at convergence. Our research question is further important because GNNs with a PA bias can amplify degree disparities, which translates to increased discrimination and disparities in social influence among nodes.

As we uncover this new form of unfairness, there are no existing solutions to this unfairness in the literature. We propose a training-time regularization-based fairness method that alleviates this unfairness without greatly sacrificing the test AUC of LP. While capping the number of positive link predictions per node is a possible solution, doing so with utility in mind requires identifying a utility-maximizing subset of link predictions. As our theoretical and empirical results reveal, GCN LP scores are often inherently proportional to the geometric mean of the degrees of the incident nodes, which can make them a poor indicator of prediction confidence; from a calibration perspective, GCNs naturally make overconfident predictions for links between high-degree nodes.

While we describe methods for alleviating degree bias in §2, these methods address degraded performance for low-degree nodes, not PA bias. We do not study performance issues but rather how GCNs scale representations of nodes proportionally to (approximately) the square root of their within-group degree, which affects the magnitude of their LP scores (cf. §K).

In summary, we augment the field's understanding of degree bias beyond performance disparities across nodes. We further lay a foundation to study PA bias and within-group unfairness in GNN LP more broadly (e.g., SOTA contrastive methods for LP), which is a critical and interesting direction of research.

[7] https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.link_prediction.preferential_attachment.html

K. Comparison to Prior Research on Degree Bias

Studies concerning degree bias have observed that low-degree nodes experience degraded performance compared to high-degree nodes. They have thus often formulated degree bias from a performance perspective, focusing on equal opportunity. In particular, these studies seek to satisfy $\mathbb{P}(\hat{y}_v = y | y_v = y, deg(v) = d) = \mathbb{P}(\hat{y}_v = y | y_v = y, deg(v) = d')$ for all possible degrees d, d', where $\hat{y}_v$ is the prediction for node v and y_v is its ground-truth label. This fairness criterion treats the degree of a node as a sensitive attribute, requiring that a GNN's accuracy is consistent across nodes with different degrees.

However, in this paper, we seek to ensure that degree disparities in networks are not amplified by GNN LP. We cannot adopt the equal opportunity formulation of degree bias because it is concerned with performance while we are concerned with degree disparity amplification. For example, even if we consistently predict links with the same accuracy across nodes with different degrees, high-degree nodes can still receive higher LP scores than low-degree nodes. In this way, the "degree bias" discussed by other studies is not compatible with our unfairness metric (Eqn. 17). We also cannot simply adopt common LP fairness metrics like dyadic fairness, as they do not capture the new type of unfairness that we uncover.

Roughly, we care that $\mathbb{E}[\hat{y}_{uv} | deg(u) = d] = \mathbb{E}[\hat{y}_{uv} | deg(u) = d']$, where $\hat{y}_{uv}$ is the GNN score for a link prediction between nodes u, v. In other words, we do not want GNN LP scores to be higher for high-degree nodes vs. low-degree nodes. This is what motivates our fairness metric (Eqn. 17).

Our theoretical analysis (Theorem 4.3) and empirical validation (§6.1) reveal that GCNs fundamentally often predict links between nodes i, j with score approximately $\propto \sqrt{deg(i) \cdot deg(j)}$ *because* of their symmetric normalized filter. This finding of a preferential attachment bias allows us to express our unfairness metric in terms of degree disparity (Eqn. 22), but this degree disparity is *not* related to the "degree bias" that has been discussed by other papers; this is a new fairness paradigm.

L. Justification of Assumptions in Lemma 4.1

The independence of path activation probabilities may not always hold true in practice. However, we verify that this assumption is plausible via our extensive experiments on real-world datasets that validate our theoretical analysis (§6.1). This assumption also aligns with findings that deep neural networks have an inductive bias towards learning simpler, often linear, functions (Nakkiran et al., 2019; Valle-Pérez et al., 2019). Furthermore, a variant of our assumption (where $\rho(i) = \rho$ is constant for all nodes) has been used in the literature to simplify theoretical analysis (e.g., Xu et al. (2018); Tang et al. (2020)); our assumption may be more realistic than this variant, as it captures that the probability of paths activating can differ across nodes (e.g., due to differences in features, neighborhood structure).