Toward Understanding Mechanisms of Unfairness and Moving Beyond Demographic Attributes

David M. Liu
Northeastern University Khoury College of Computer Sciences

MSR New England ML Ideas Seminar
February 20, 2024

dliu18.github.io

liu.davi@northeastern.edu

@dayvidliu

Northeastern University

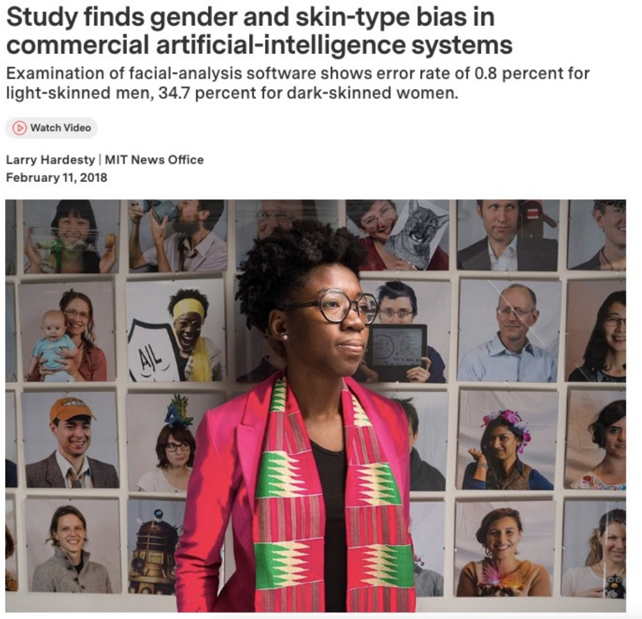
Unfairness of ML for Decision Making

Algorithms Allegedly Penalized Black Renters. The US Government Is Watching

The Department of Justice warned a provider of tenant-screening software that its technology must comply with fair housing law.

ILLUSTRATION: JACQUI VANLIEW; GETTY IMAGES

Source: Wired



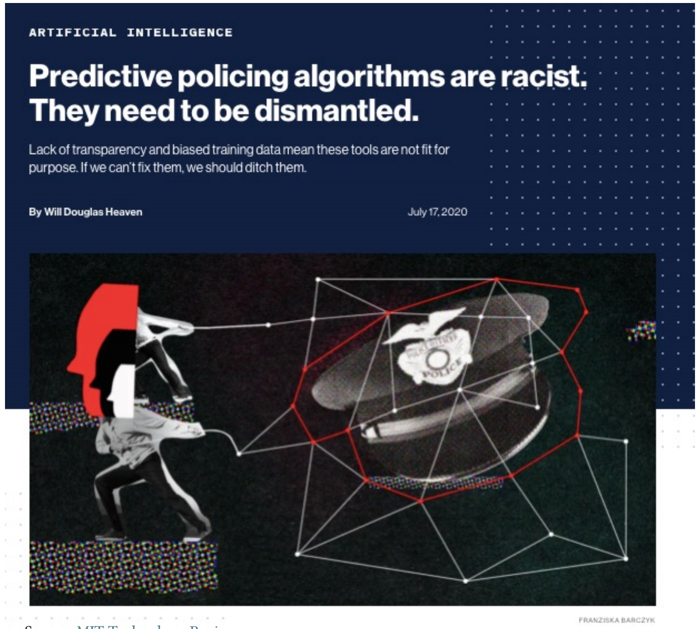
Study finds gender and skin-type bias in commercial artificial-intelligence systems

Examination of facial-analysis software shows error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women.

Watch Video

Larry Hardesty | MIT News Office
February 11, 2018

Source: MIT News



ARTIFICIAL INTELLIGENCE

Predictive policing algorithms are racist. They need to be dismantled.

Lack of transparency and biased training data mean these tools are not fit for purpose. If we can't fix them, we should ditch them.

By Will Douglas Heaven July 17, 2020

FRANZISKA BARCZYK

Source: MIT Technology Review

Defining Group Fairness

Goal: balance classifier performance across sensitive-attribute groups

Statistical Parity – Corbett-Davies et al. (2017)

$$P\left(\hat{Y} = 1 \,\middle|\, A = 0\right) = P\left(\hat{Y} = 1 \,\middle|\, A = 1\right)$$

Equalized Odds – Hardt, Price, Srebro (2016)

$$P\left(\hat{Y} = 1 \,\middle|\, A = 0, y = 1\right) = P\left(\hat{Y} = 1 \,\middle|\, A = 1, y = 1\right)$$

Calibration – Chouldechova (2017)

$$P(Y = 1 \mid A = 0, S = s) = P(Y = 1 \mid A = 1, S = s)$$

Symbol	Meaning
Y	Ground truth label
$\hat{Y}$	Predicted label
A	Sensitive attribute
S	Risk score

Limitations of Existing Group Fairness Approaches

1. Rely on sensitive/demographic attributes

2. Don’t help us understand sources of *model* unfairness in the first place

Overview

Tackling the limitations:

1. Rely on demographic attributes
   - ➢ Defining group fairness with social networks [FAccT ‘23]
2. Don’t help us understand sources of unfairness
   - ➢ Identify mechanisms of unfairness in PCA collaborative filtering [In Submission]

Moving Beyond Demographic Attributes

Group fairness without demographics using social networks

FAccT'23

David Liu
Northeastern

Virginie Do
Meta AI

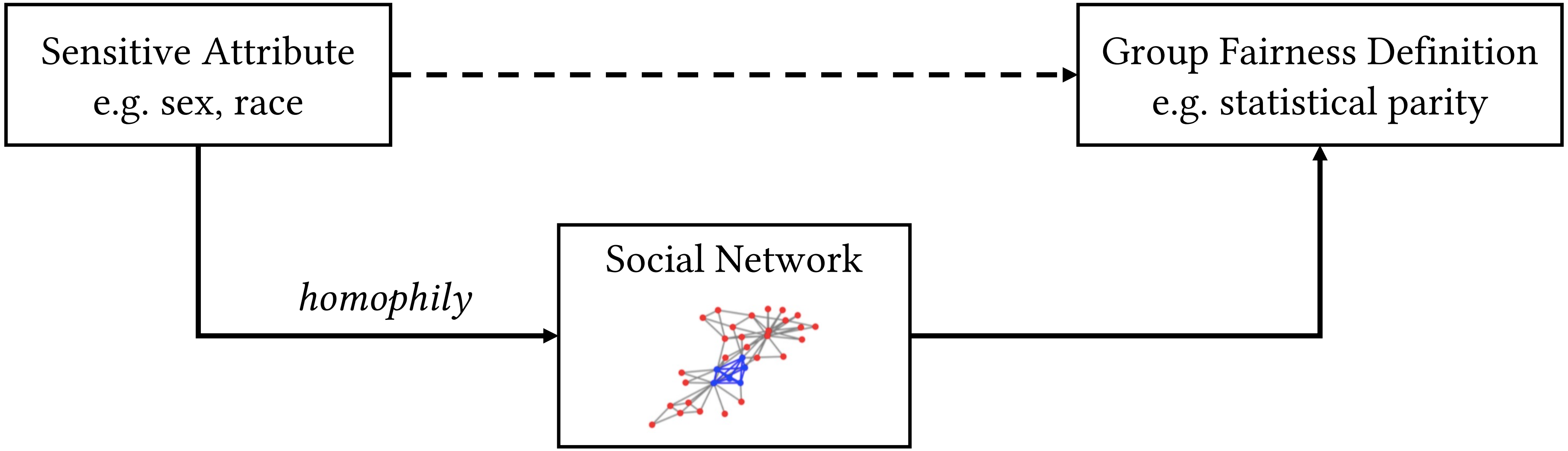
Nicolas Usunier
Meta AI

Max Nickel
Meta AI

Fairness Without Demographics

Group fairness definitions traditionally rely on sensitive attribute labels to define groups of individuals.

However, often these labels are unavailable or harmful to collect.



Sensitive Attribute
e.g. sex, race

Group Fairness Definition
e.g. statistical parity

homophily

Social Network

Question: if we instead have access to a social network, it is possible to measure group fairness without assigning group labels in the process?

Our Contributions

1. **[Social Network Homophily]** We propose a novel measure of group fairness that does not depend on group labels and instead uses homophily in social networks to reduce inequality in outcomes.

2. **[Group-Free Group Fairness]** Our approach is a measure of inequality that is "group-free" in that it avoids attempting to define groups entirely and is solely based on the similarities of individuals.

3. **[Evaluation]** We theoretically analyze our measure of group-free group fairness and empirically evaluate it on three tasks: classification, maximizing information access, and recommendation.

Review of Homophily

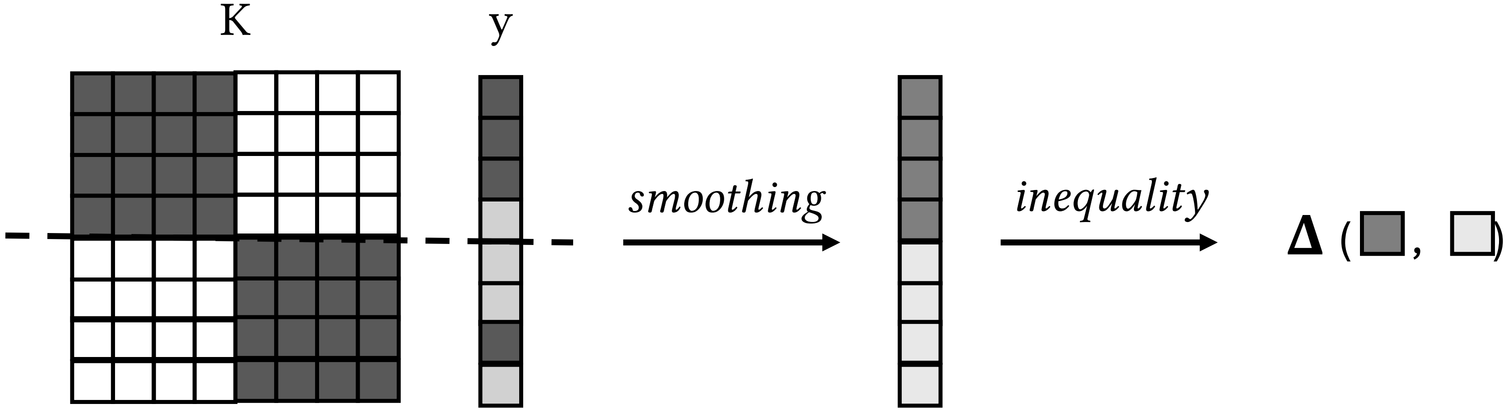
- The tendency of individuals to connect at a higher rate with people that share similar characteristics. *“Birds of a feather flock together”*.
- Widely observed across many attributes and types of connections.
  - Race
  - Sex
  - Age
  - Religion
  - Education, occupation, social class

Group-free Group Fairness

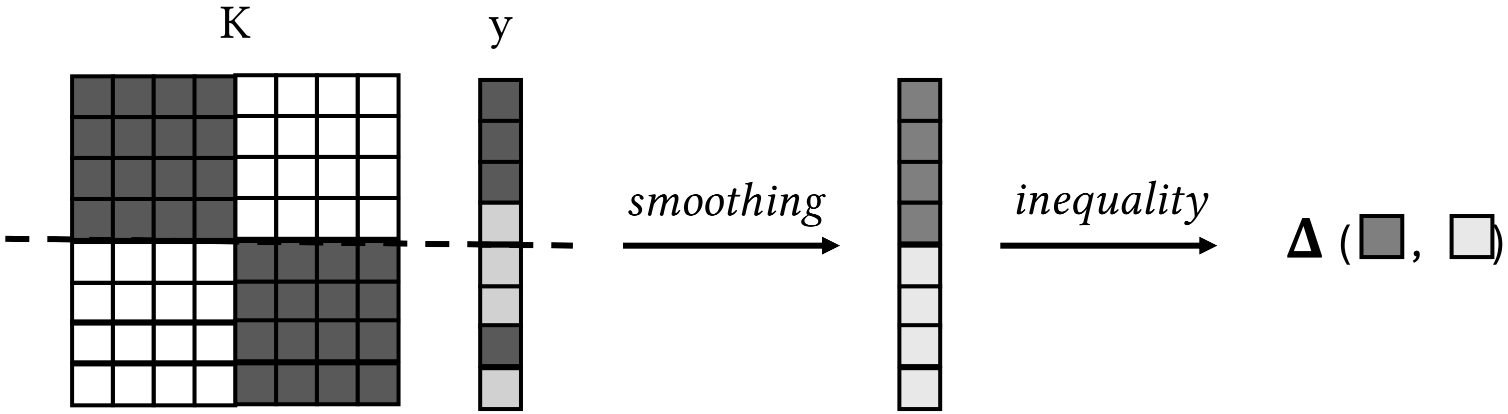
Notation: Let $K \in \mathbb{R}^{n \times n}$ be a kernel where K_{ij} is the similarity between i and j. And, let y be a vector of individual outcomes.

Our approach, group-free group fairness, involves two functions:

1. Smoothing function
2. Economic inequality function



Group-free Group Fairness



Our measure of group-free group fairness (Δ_b) is defined as:

$$\Delta_b = F(A(K, y), K1)$$

➢Function A smooths the outcomes within groups e.g. $A(K, y) = \frac{Ky}{K1}$

➢Function F is an economic inequality measure (e.g. normalized variance)

Inferring Kernels from Networks

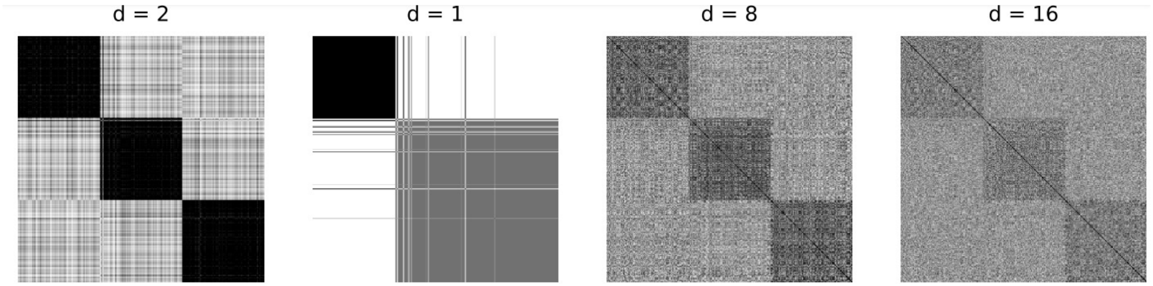
1. Embed the nodes as $X \in \mathbb{R}^{n \times d}$ (we use Laplacian Eigenmaps)

2. Define a similarity matrix S where entry S_{ij} is the cosine similarity between X_i and X_j.

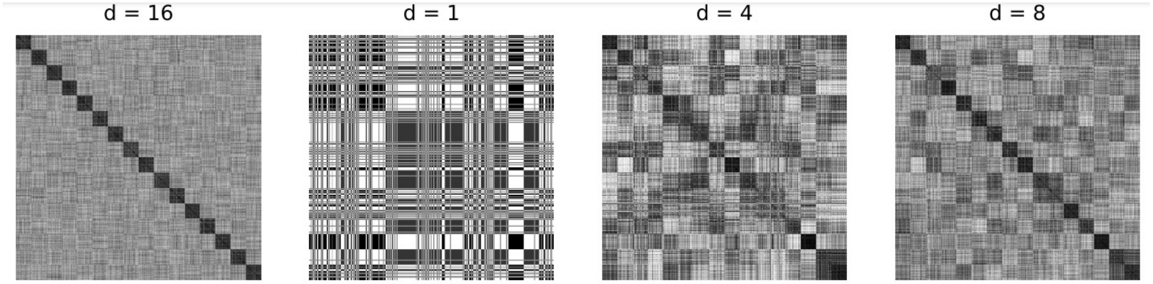
3. Let K be S following *column* normalization.

Stochastic Block Model Examples

3 blocks



16 blocks



Analysis: Additive Decomposability

Given group labels, an inequality function F satisfies additive decomposability if:

$$F(x) = \Delta_{within}(x) + \Delta_{between}(x)$$

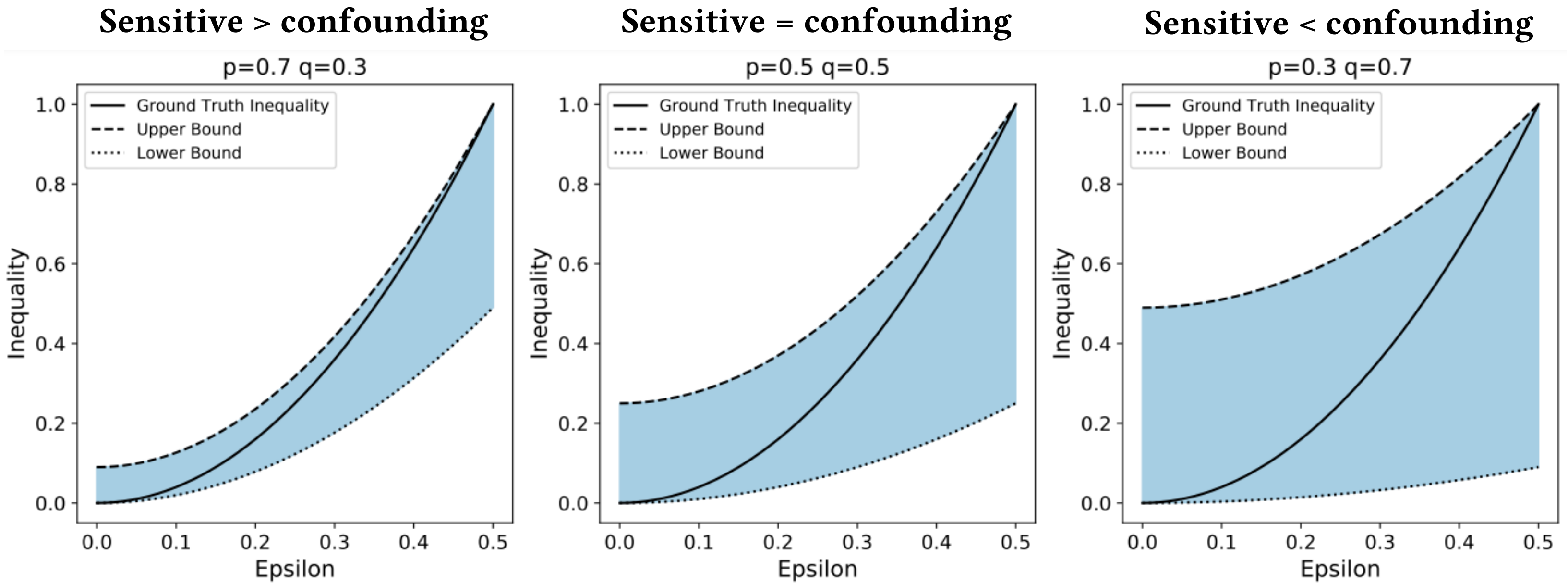
We port the property to the pairwise similarity setting. i.e. an measure of inequality satisfies additive decomposability if and only if:

$$F(\mathbf{y}) = \left[\sum_{i=1}^{n} \frac{q(A(K,\mathbf{y})_i)K_i^{\top}\mathbf{1}}{q(\mu)n} F\left(\mathbf{y}, K_i\right)\right] + \underbrace{F\left(A\left(K,\mathbf{y}\right), K\mathbf{1}\right)}_{\Delta_b}$$

Analysis: Bounding Confounding Features

Observe that there may be homophilous attributes that are not sensitive.

We analyze a simple setting involving two groups of equal size where each node is labeled 0 or 1. Let p be the strength of the sensitive attribute, q be the strength of the confounding attribute, and ϵ is the difference in group outcome averages.



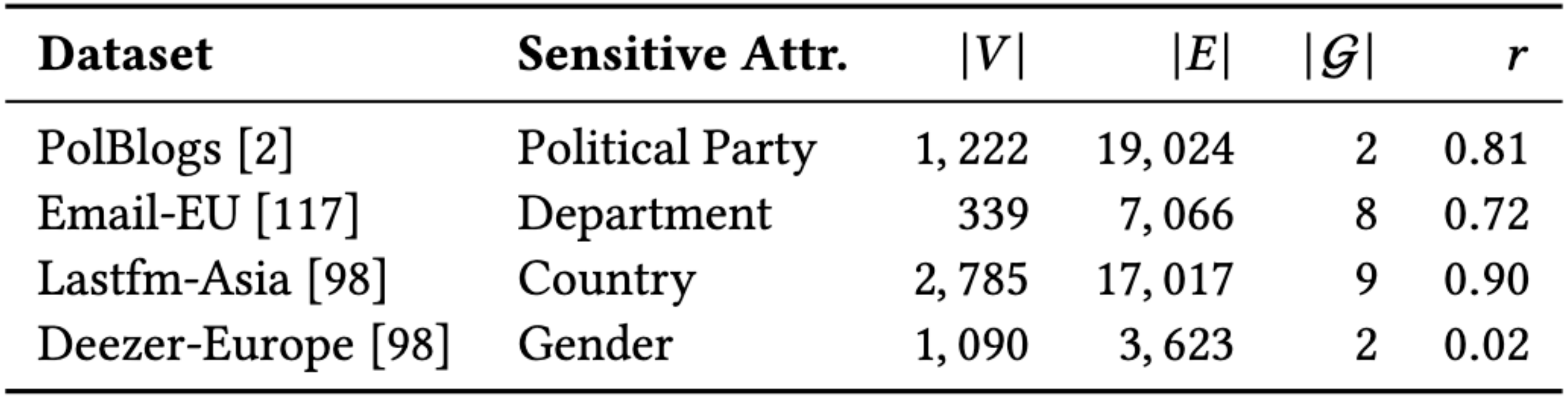
Sensitive > confounding **Sensitive = confounding** **Sensitive < confounding**

The black line is the ground-truth value and our measure returns a value in the shaded region.

Experimental Setup

Across all evaluation tasks, we use datasets that provide a network and ground-truth sensitive attribute labels.

The labels are used *only during test time* to evaluate the kernel-based approach.



Dataset	**Sensitive Attr.**	$\lvert V \rvert$	$\lvert E \rvert$	$\lvert \mathcal{G} \rvert$	r
PolBlogs [2]	Political Party	$1,222$	$19,024$	2	0.81
Email-EU [117]	Department	339	$7,066$	8	0.72
Lastfm-Asia [98]	Country	$2,785$	$17,017$	9	0.90
Deezer-Europe [98]	Gender	$1,090$	$3,623$	2	0.02

Compare against a community detection baseline which assigns discrete labels.

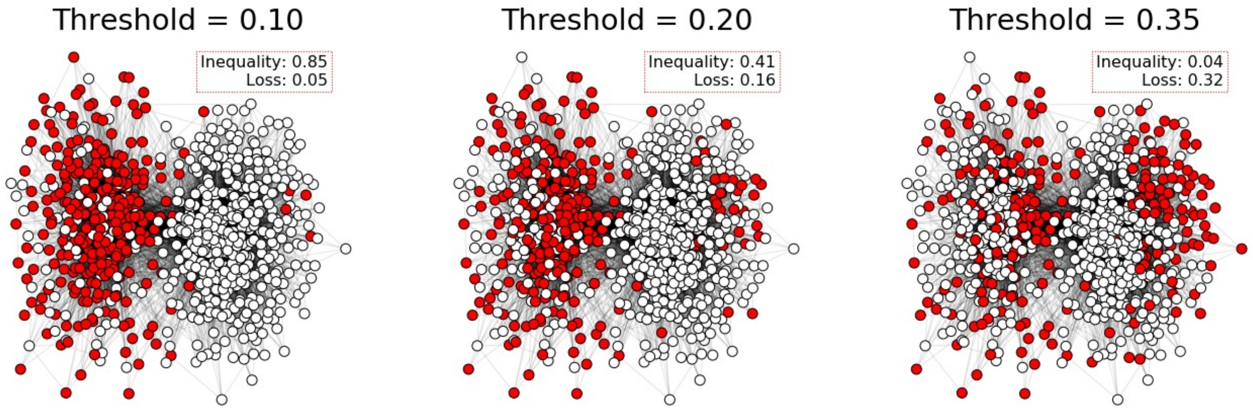
* Graph statistics are following pre-processing, which is detailed in the full paper.

Evaluation: Node Classification

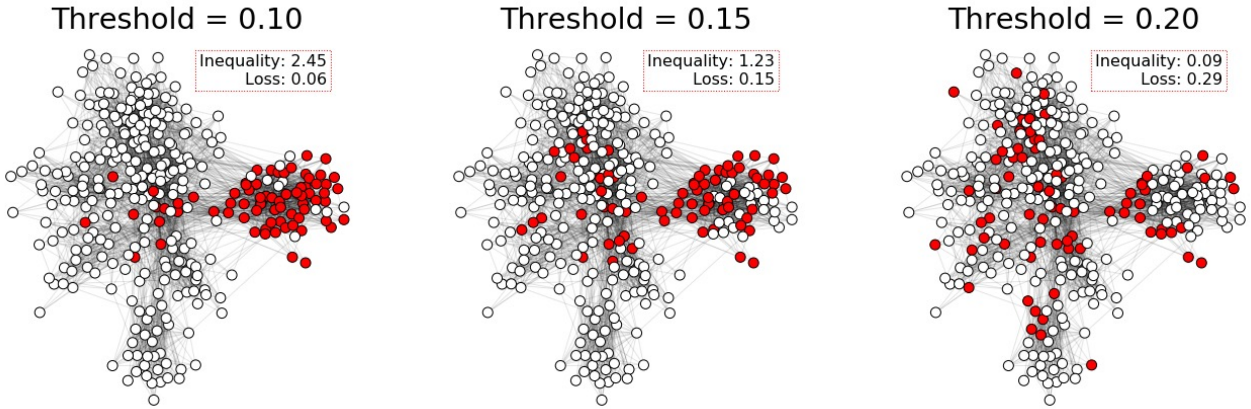
Setup: a classifier labels nodes 0 or 1 (red), where 1 is the desired outcome.

We post-process the labels and relabel nodes such that each smoothed value is above a minimum threshold:
$A(K, y) \geq \theta_{min}$

PolBlogs



Email-Eu

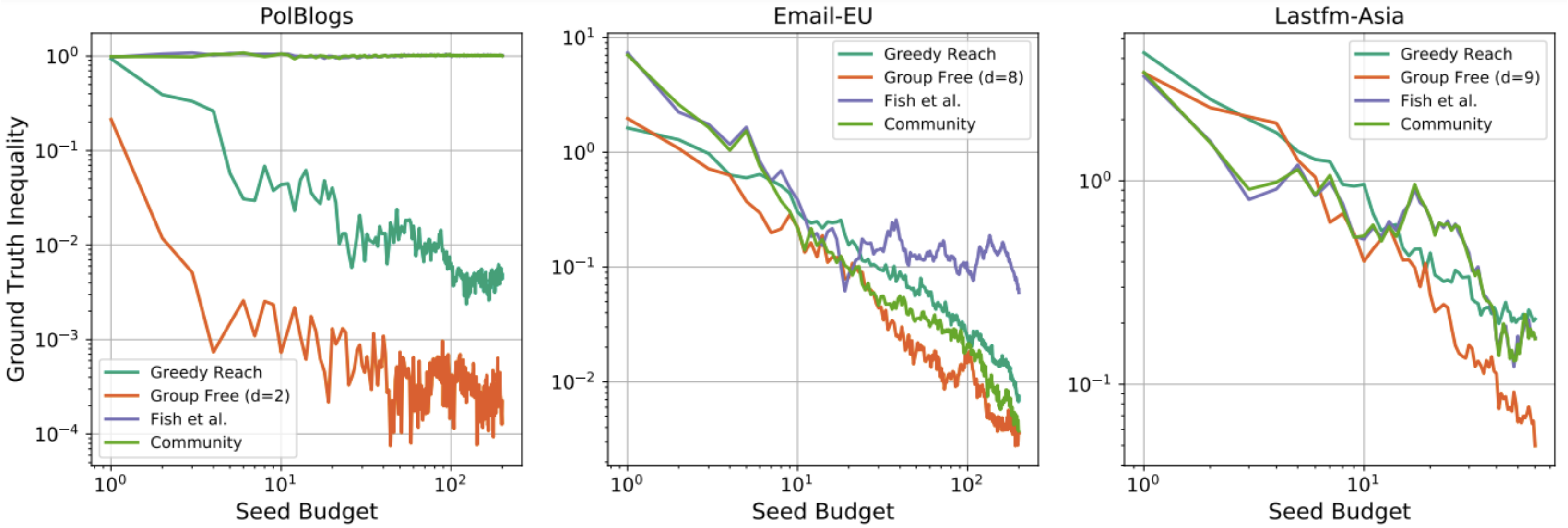


Evaluation: Maximizing Information Access

Goal: maximize the reach of a helpful piece of information (e.g. vaccine access) cascading in a network. The task is to choose the initial seed node set.

We use an algorithm (orange) *that greedily selects the node that maximizes* $min\ A(K, y)$.

Compare against baselines that greedily maximize total reach, min individual exposure, min community exposure. Y-axis is inequality measured with labels; lower is better.

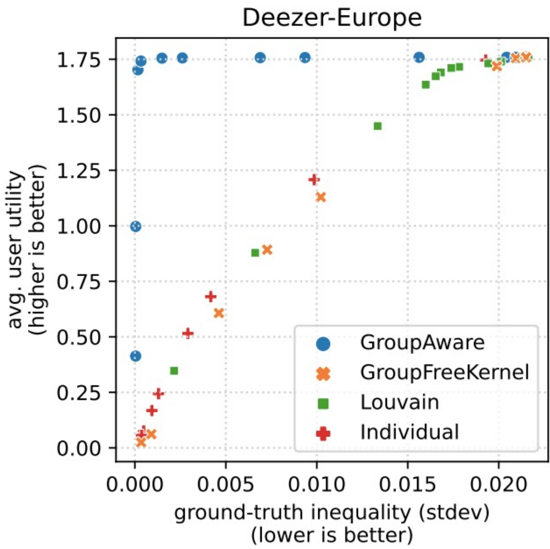


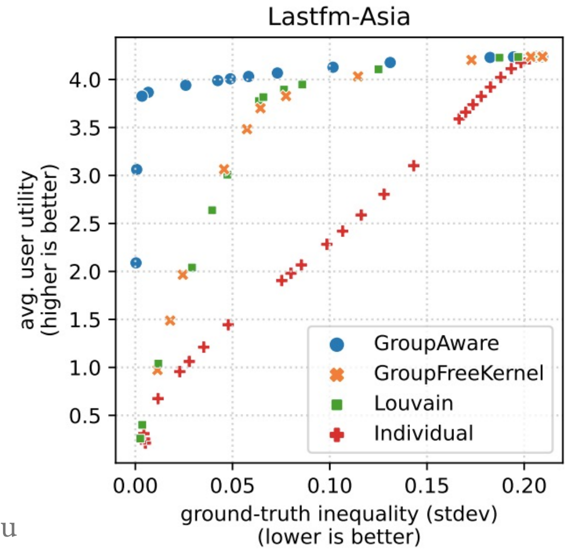
Evaluation: Recommender systems

Goal: given a set of items and each user’s preference ratings over the items, provide a ranking of items to each user that *balances user utility and as well as group fairness for users (each item should be exposed to all groups of users).*

Optimize a joint objective that rewards user utility and penalizes unequal item exposure.

Our group-free approach outperforms the baseline that treats each item individually but does not perform as well as knowing the group labels.





Group-Free Group Fairness: Limitations and Takeaways

Limitations

1. Network datasets can be biased in how they are collected and sampled.
2. Our method requires discernable community structure.
3. Assumes homophilous network formation.
4. Embedding process may require tuning (choice of algorithm, number of dimensions)

Takeaways

1. We present a measure for group-free group fairness that is based on similarities between individuals.
2. Our evaluations show that our measure can reduce inequality among groups given only the network.
3. In the process, we do not infer the sensitive attribute or assign any group labels.

Overview

Tackling the limitations:

1. Rely on demographic attributes
   - ➢ Defining group fairness with social networks [FAccT ‘23]

2. Don’t help us understand sources of unfairness
   - ➢ Identify mechanisms of unfairness in PCA collaborative filtering [In Submission]

Toward Understanding Unfairness Mechanisms

When Collaborative Filtering is not Collaborative: Unfairness of PCA for Recommendations

In submission

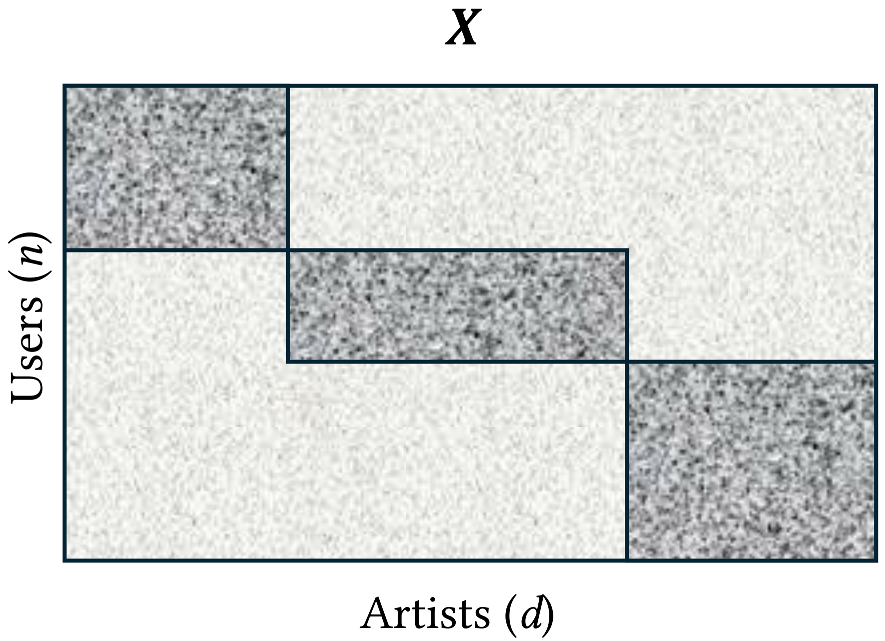
David Liu
Northeastern

Jackie Baek
NYU Stern

Tina Eliassi-Rad
Northeastern

A Motivating Example

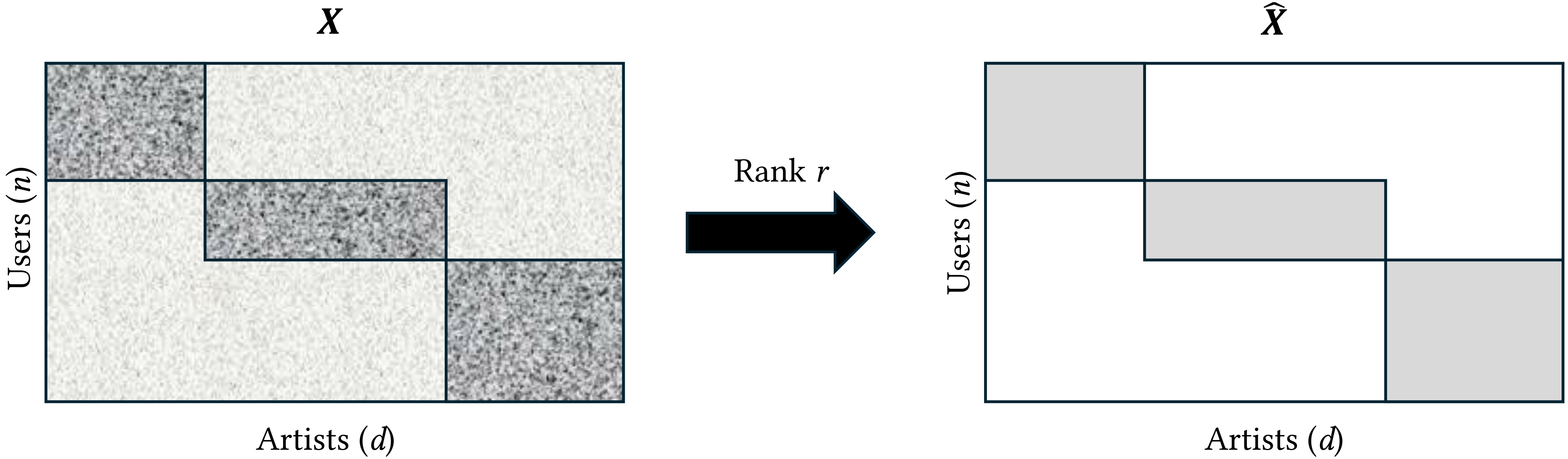
Last.fm music listening dataset of user listening counts (Cantador et al., 2011)



A Motivating Example

Last.fm music listening dataset of user listening counts (Cantador et al., 2011)

The promise of low rank: with a few latent dimensions you can well approximate a high-dimensional matrix.



A Motivating Example

Globally, as the rank budget *r* increases,
$|X - \hat{X}|^2$ is (exponentially) decreasing.

However, are all portions of the matrix equally well approximated?

Let *Normalized Item Error* be the artist-level error as r increases:

$$\text{Normalized Item Error} = \frac{\left|X_{.j} - \hat{X}_{.j}\right|^2}{\left|X_{.j}\right|^2}$$

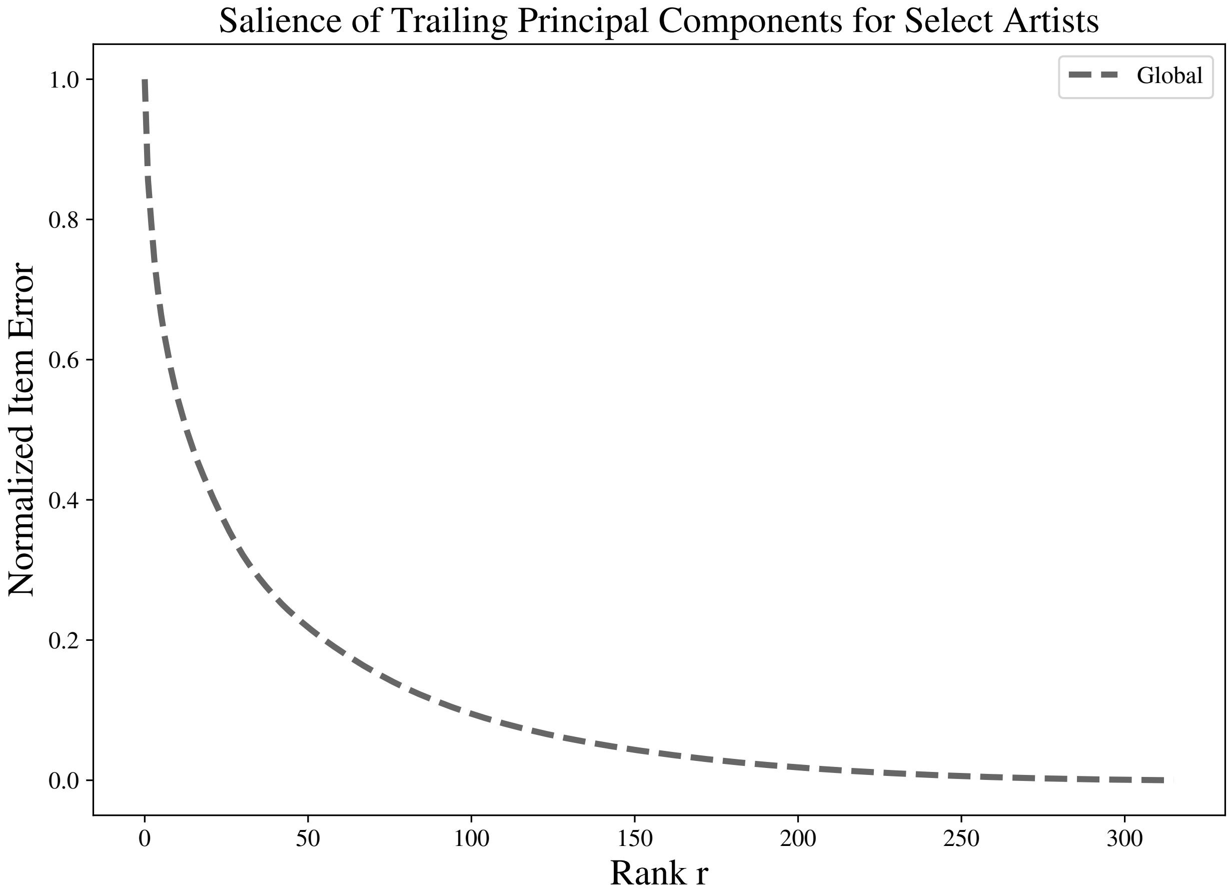
A Motivating Example

Globally, as the rank budget r increases, $|X - \hat{X}|^2$ is (exponentially) decreasing.

However, are all portions of the matrix equally well approximated?

Let *Normalized Item Error* be the artist-level error as r increases:

$$\text{Normalized Item Error} = \frac{\left|X_{.j} - \hat{X}_{.j}\right|^2}{\left|X_{.j}\right|^2}$$



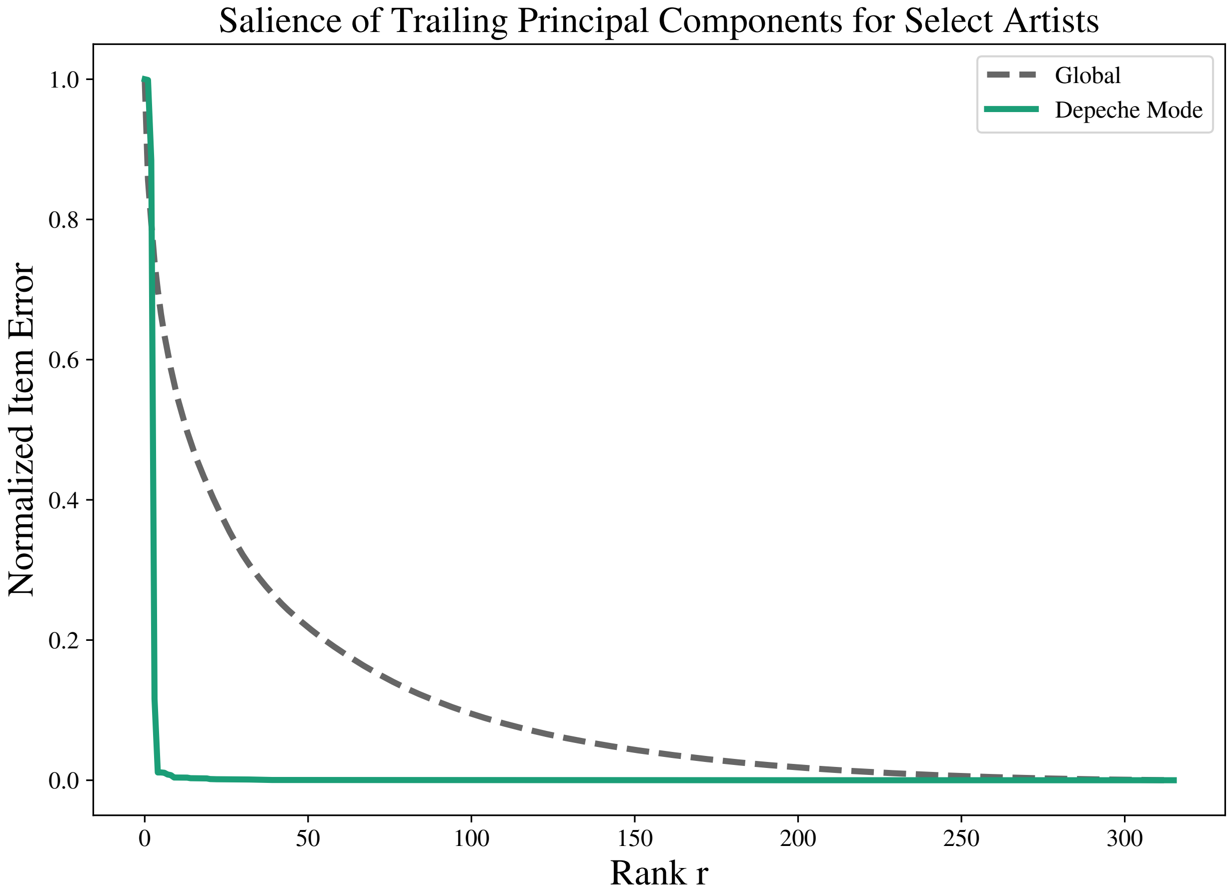
A Motivating Example

Globally, as the rank budget r increases, $|X \; - \hat{X}|^2$ is (exponentially) decreasing.

However, are all portions of the matrix equally well approximated?

Let *Normalized Item Error* be the artist-level error as r increases:

$$\text{Normalized Item Error} = \frac{\left|X_{.j} - \hat{X}_{.j}\right|^2}{\left|X_{.j}\right|^2}$$



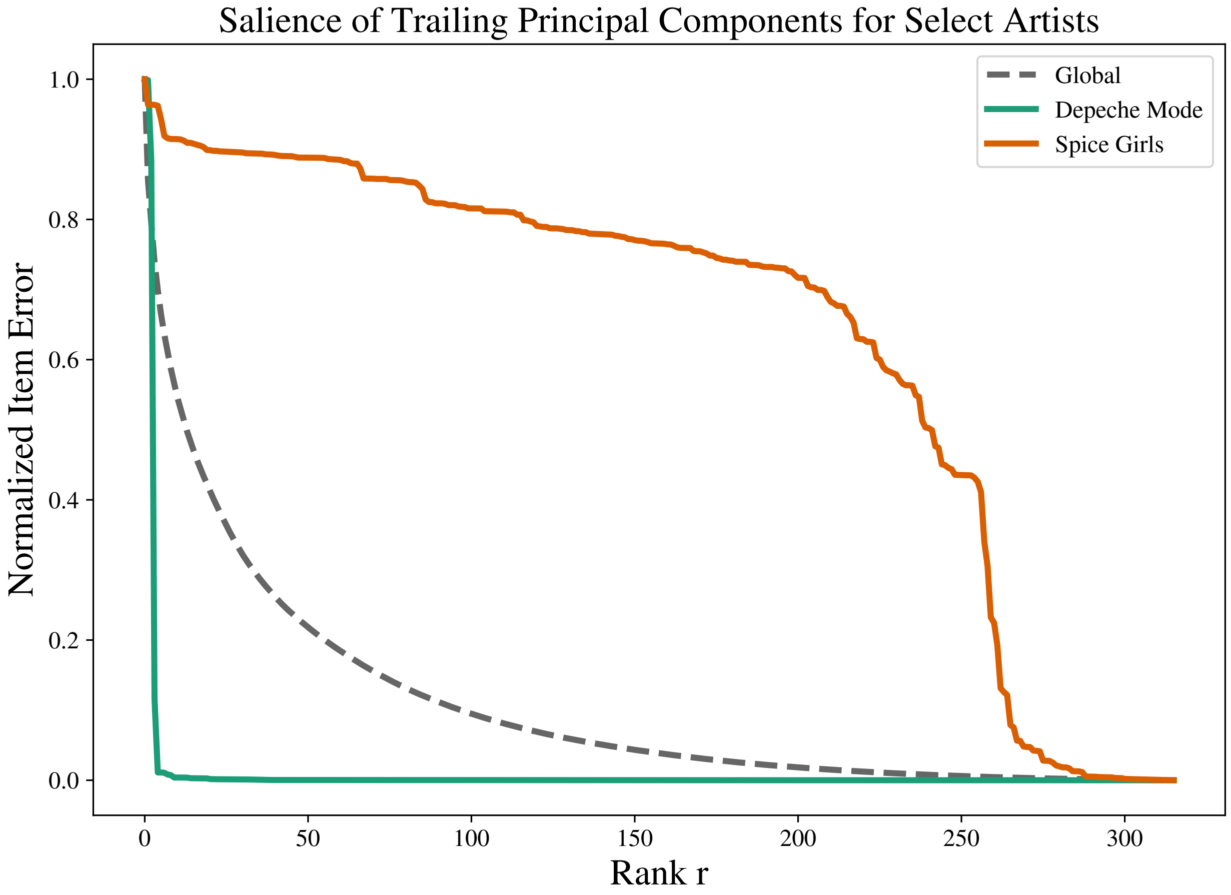
A Motivating Example

Globally, as the rank budget r increases, $|X - \hat{X}|^2$ is (exponentially) decreasing.

However, are all portions of the matrix equally well approximated?

Let *Normalized Item Error* be the artist-level error as r increases:

$$\text{Normalized Item Error} = \frac{\left|X_{.j} - \hat{X}_{.j}\right|^2}{\left|X_{.j}\right|^2}$$



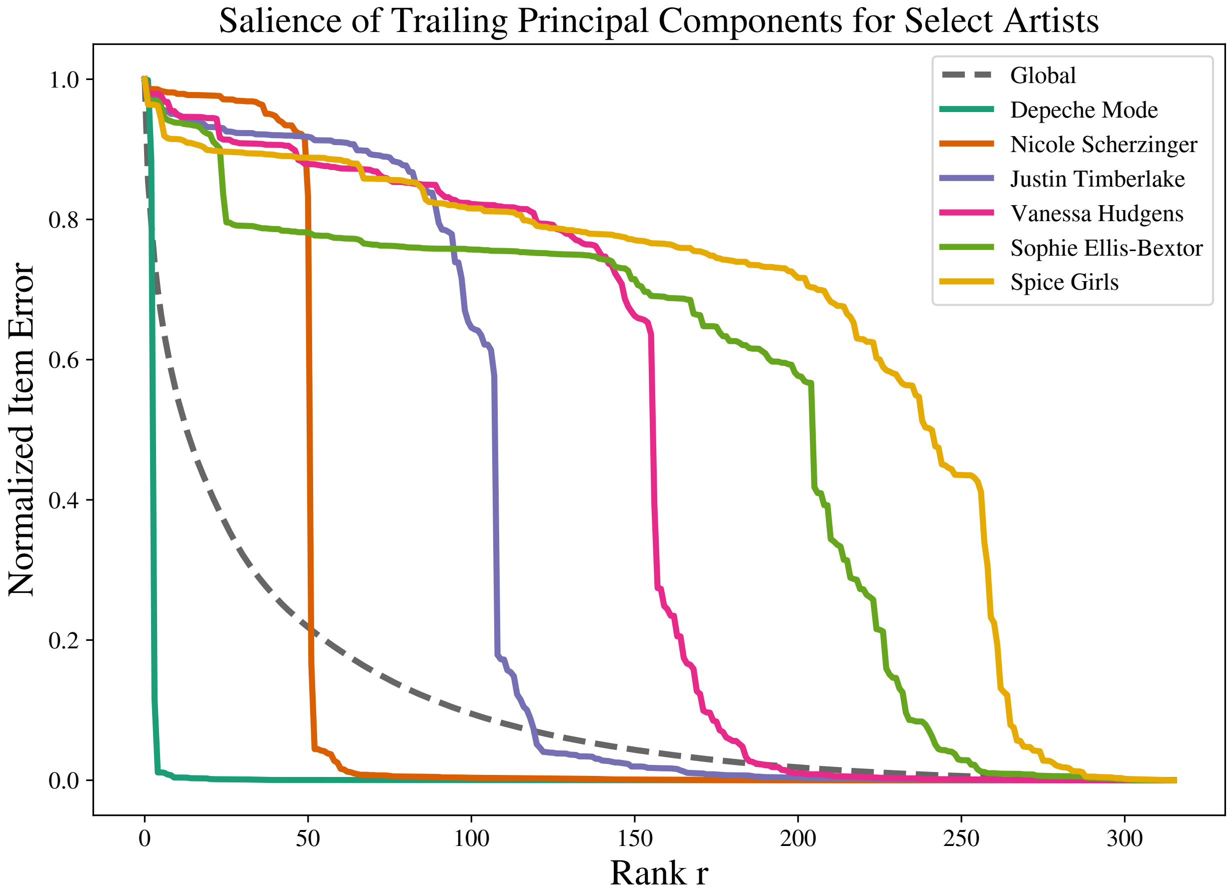
A Motivating Example

Globally, as the rank budget *r* increases, $|X - \hat{X}|^2$ is (exponentially) decreasing.

However, are all portions of the matrix equally well approximated?

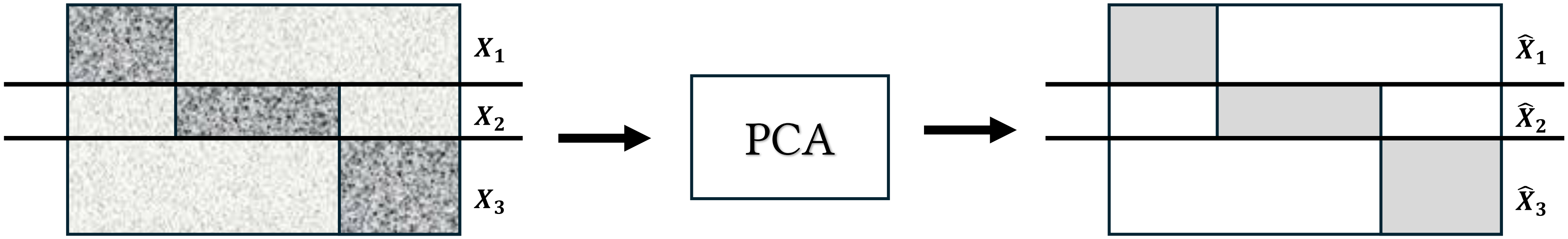
Let *Normalized Item Error* be the artist-level error as r increases:

$$\text{Normalized Item Error} = \frac{|X_{.j} - \hat{X}_{.j}|^2}{|X_{.j}|^2}$$



From Fairness Definitions to Mechanisms

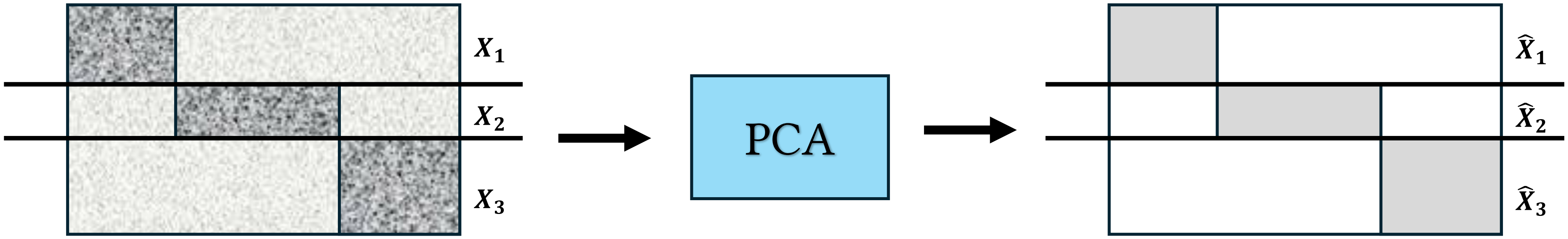
Prior work on fair PCA



$$f\left(\left|X_1 - \hat{X}_1\right|^2, \dots, \left|X_g - \hat{X}_g\right|^2\right)$$

From Fairness Definitions to Mechanisms

Prior work on fair PCA



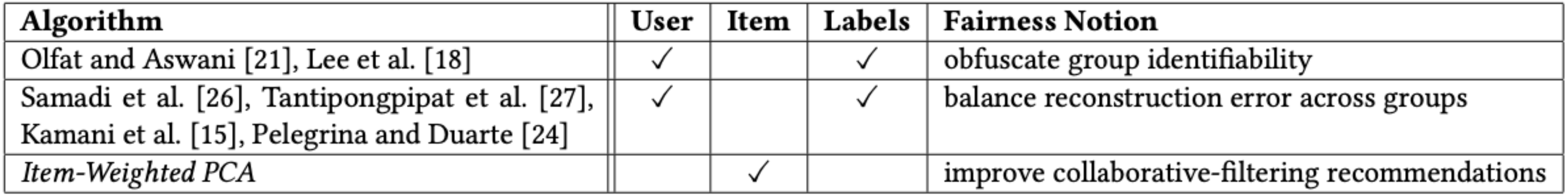
$$f(|X_1 - \hat{X}_1|^2, \dots, |X_g - \hat{X}_g|^2)$$

RQ: What are the mechanisms of unfairness in PCA in the first place?

From Fairness Definitions to Mechanisms

We identify mechanisms stemming from disparities in item popularities. Leading us to look at:

- Item over user disparities
- Implications for recommender systems



Algorithm	**User**	**Item**	**Labels**	**Fairness Notion**
Olfat and Aswani [21], Lee et al. [18]	✓		✓	obfuscate group identifiability
Samadi et al. [26], Tantipongpipat et al. [27], Kamani et al. [15], Pelegrina and Duarte [24]	✓		✓	balance reconstruction error across groups
Item-Weighted PCA		✓		improve collaborative-filtering recommendations

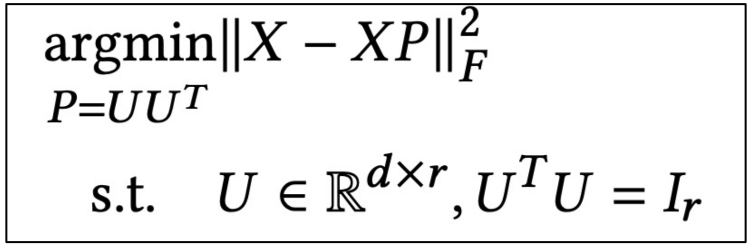
Outline

- **Mechanisms of unfairness in PCA**
    - Mechanism 1: unfairness for unpopular items
    - Mechanism 2: unfairness for popular items
- ***Item Weighted PCA*: an item re-weighting framework algorithm**
    - Efficient algorithm for improving representations of unpopular items
    - Optimality in stylized setting
- **Recommender system evaluation**
    - Improved recommendations for both popular and unpopular items

Mechanisms of Unfairness in PCA

Recap of PCA and Collaborative Filtering

PCA refresher: identifying r basis vectors to project data



$$\underset{P=UU^T}{\operatorname{argmin}} \|X - XP\|_F^2$$
$$\text{s.t.} \quad U \in \mathbb{R}^{d\times r}, U^T U = I_r$$

Recap of PCA and Collaborative Filtering

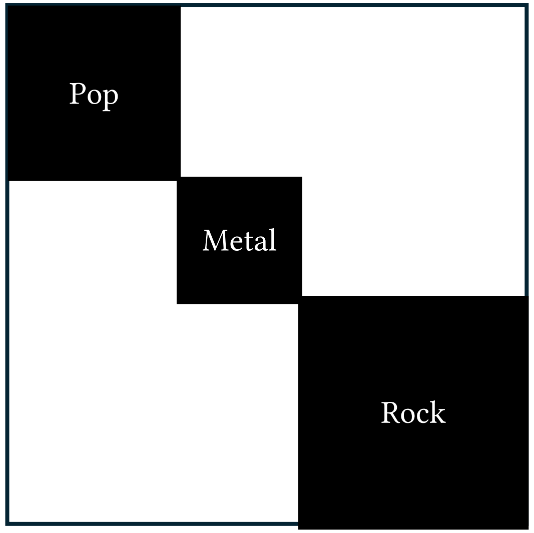
PCA refresher: identifying r basis vectors to project data

$$\boxed{\begin{aligned}&\underset{P=UU^T}{\operatorname{argmin}} \|X - XP\|_F^2 \\ &\text{s.t.} \quad U \in \mathbb{R}^{d\times r}, U^T U = I_r\end{aligned}}$$

Collaborative filtering: leverage similarities among items to infer missing values in X.

Think of P as a d x d matrix of item similarities

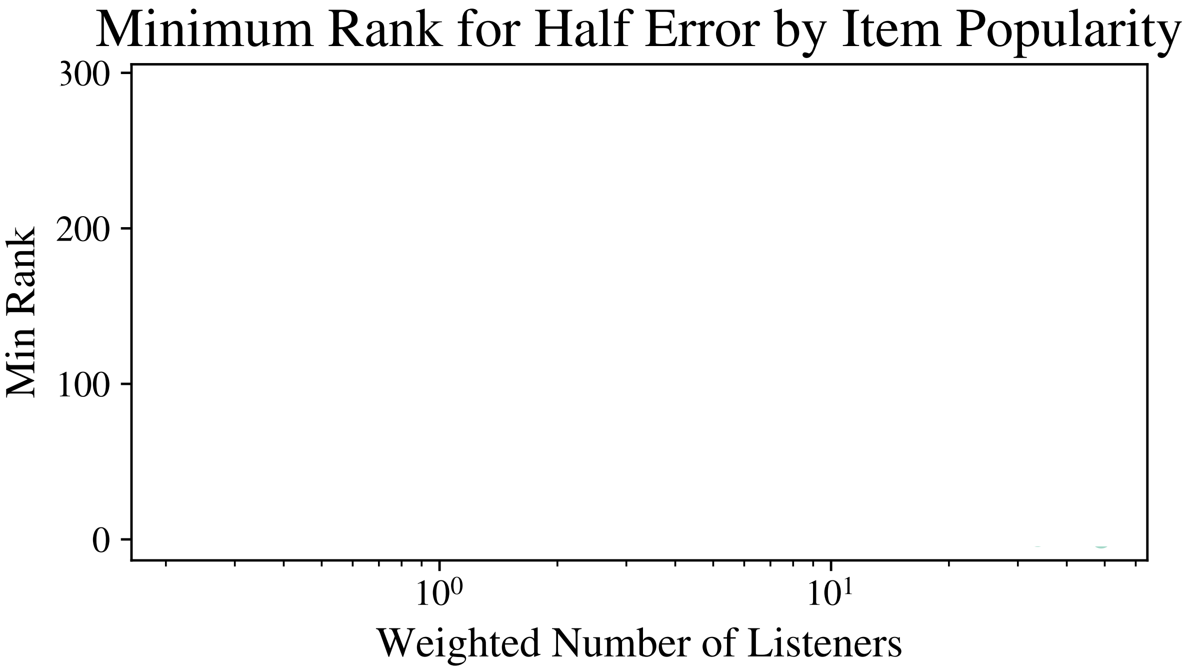
$$\hat{X}_{ij} = \sum_{j'} P_{jj'} X_{ij'}$$



Mechanism 1: Unpopular Items

The leading principal components disproportionately reconstruct entries for popular items.

$$argmin_r \frac{\left|X_{.j} - \hat{X}_{.j}\right|^2}{\left|X_{.j}\right|^2} < \frac{1}{2}$$

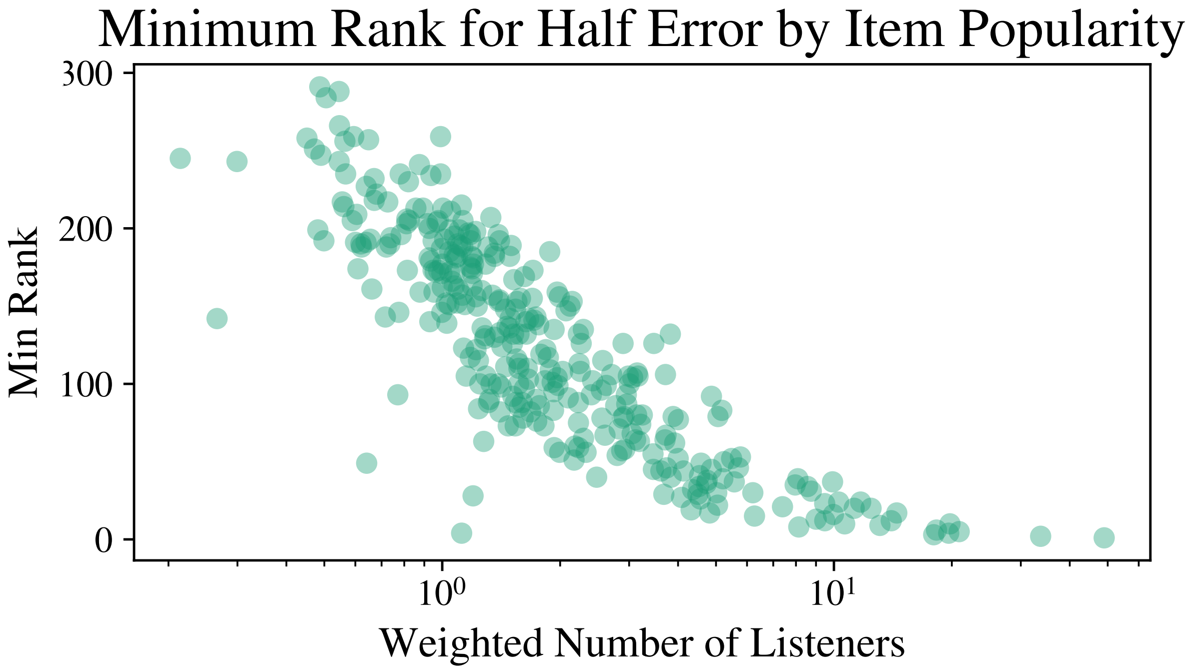


popularity

Mechanism 1: Unpopular Items

The leading principal components disproportionately reconstruct entries for popular items.

$$argmin_r \frac{\left|X_{.j} - \hat{X}_{.j}\right|^2}{\left|X_{.j}\right|^2} < \frac{1}{2}$$



popularity

Mechanism 2: Popular Items

Leading principal components specialize on individual artists as opposed to learning group information.

$$\hat{X}_{ij} = P_{jj}X_{ij} + \sum_{j' \neq j} P_{jj'}X_{ij'}$$

$P_{jj} \approx 1$ and $P_{jj'} \approx 0$

Specialization quantified by diagonal values of P

Mechanism 2: Popular Items

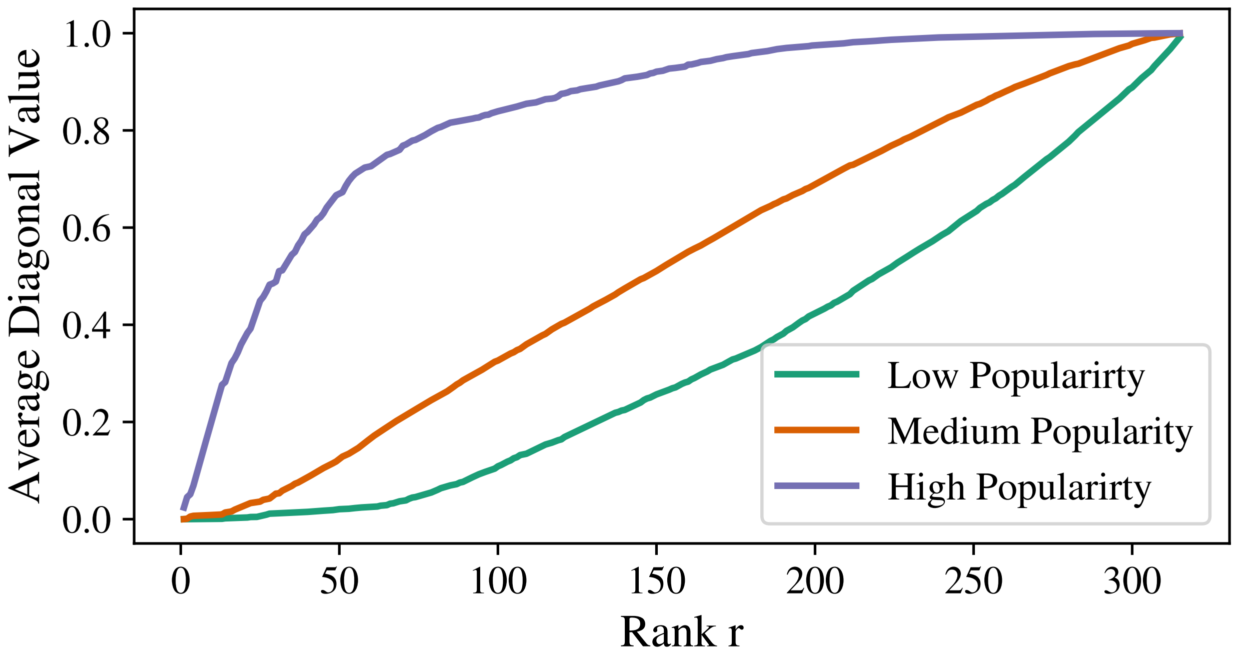
Leading principal components specialize on individual artists as opposed to learning group information.

$$\hat{X}_{ij} = P_{jj}X_{ij} + \sum_{j' \neq j} P_{jj'}X_{ij'}$$

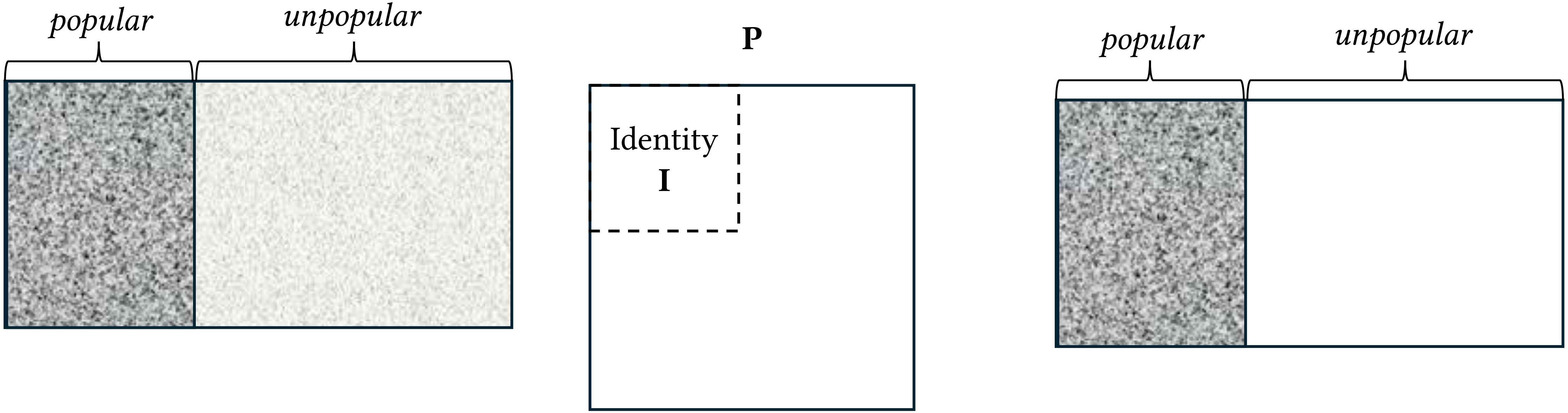
$$P_{jj} \approx 1 \text{ and } P_{jj'} \approx 0$$

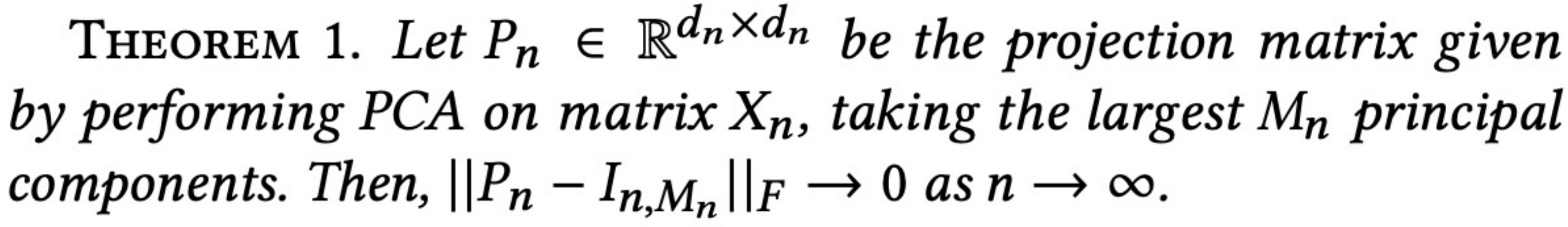
Specialization quantified by diagonal values of P

Reliance on Diagonal for Increased Rank Budget



Proven Existence of Mechanisms





Theorem 1. *Let $P_n \in \mathbb{R}^{d_n \times d_n}$ be the projection matrix given by performing PCA on matrix X_n, taking the largest M_n principal components. Then, $||P_n - I_{n,M_n}||_F \to 0$ as $n \to \infty$.*

Item-Weighted PCA

An item reweighting algorithm for improved recommendations

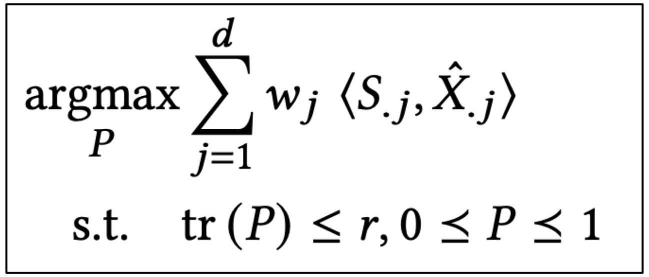
Item-Weighted PCA

Main idea: ensure that $\hat{X}$ still reflects interests in unpopular items.

$$obj = \sum_{ij} w_j \, (S_{ij} * \; \hat{X}_{ij})$$

w_j upweights less popular items

S_{ij}= sign(X_{ij}). Ensure and have the same sign.

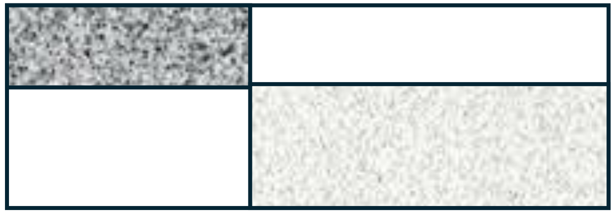


$$\begin{aligned} &\underset{P}{\operatorname{argmax}} \sum_{j=1}^{d} w_j \, \langle S_{.j}, \hat{X}_{.j} \rangle \\ &\text{s.t.} \quad \operatorname{tr}(P) \leq r, 0 \leq P \leq 1 \end{aligned}$$

Constraints are convex relaxation of rank(P) ≤ r. In paper we show this constraint is tight.

Optimality for Block Matrices

Consider binary $X \in \{0, 1\}^{n \times d}$ matrices where some items are popular and others are unpopular
Assume: each user likes only popular items or unpopular items



Theorem: Item-Weighted PCA provides the optimal solution to the popularity-normalized objective,

$$\frac{\left|X_p - \hat{X}_p\right|^2}{\left|X_p\right|^2} + \frac{\left|X_u - \hat{X}_u\right|^2}{|X_u|^2}$$

Where $w_j = \left|X_p\right|^{-1}$ for all popular items and $w_j = |X_u|^{-1}$ for all unpopular items

Baselines as Instances of Item Weighted PCA

Assume: all popular items have the same popularity ($\sum_j X_{ij} = n_p$) and all unpopular items have the same popularity ($\sum_j X_{ij} = n_u$)

Two baselines:

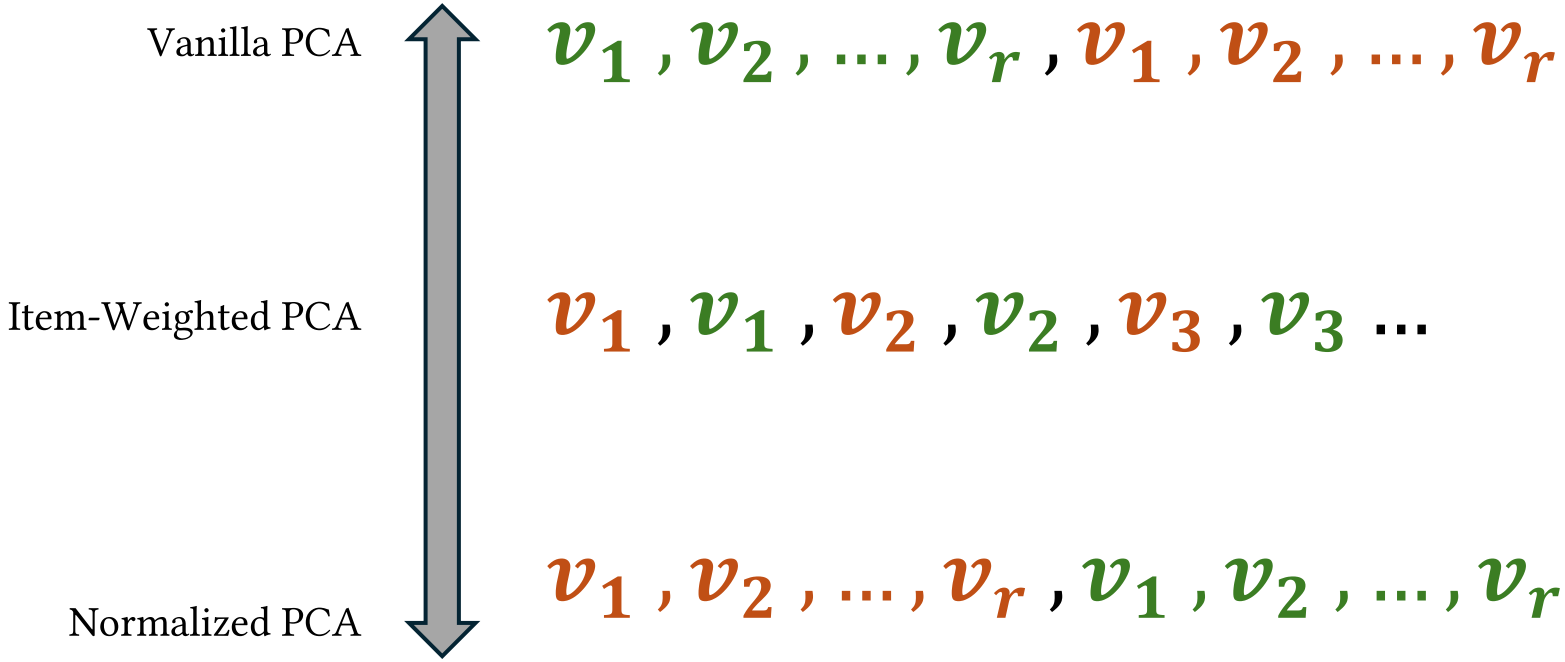
- Vanilla PCA
- Column-normalized PCA: normalize columns of X before PCA

Both baselines are instances of Item-Weighted PCA but with differing weights.

Baselines as Instances of Item Weighted PCA

If #unpopular = $\sqrt{\frac{n_p}{n_u}}$ #popular and popularity gap is sufficiently large*

Eigenvectors of X_p: $\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_r$ and X_u: $\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_r$



Vanilla PCA $\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_r, \boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_r$

Item-Weighted PCA $\boldsymbol{v}_1, \boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_3 \dots$

Normalized PCA $\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_r, \boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_r$

* Details in paper

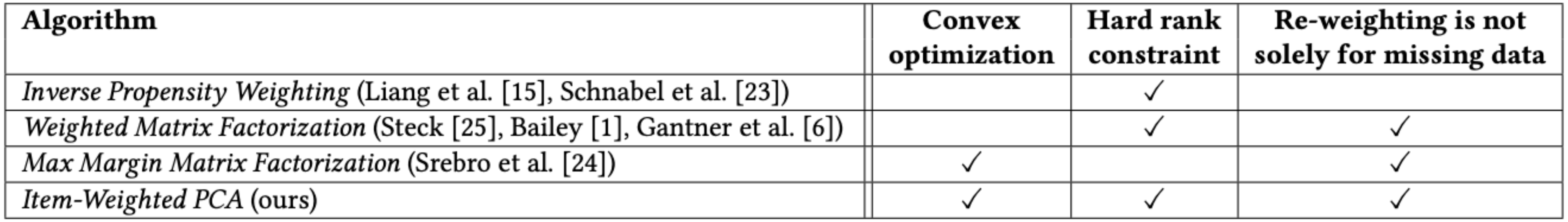
Comparison to Re-weighting Literature

Weighted Matrix Factorization (non-convex)

$$\min \sum_{ij} w_{ij}\left(r_{ij} - x_i^T y_j\right)^2 + \lambda_1 \|X\|_F^2 + \lambda_2 \|Y\|_F^2$$

For Inverse Propensity Weighting, $w_{ij} = 1/p_j$ where p_j is the *propensity* of item *j*. Accounts for the fact that data (ratings) are not missing at random.

Difficult to enforce ***convexity*** and a ***hard rank constraint***



Algorithm	**Convex optimization**	**Hard rank constraint**	**Re-weighting is not solely for missing data**
Inverse Propensity Weighting (Liang et al. [15], Schnabel et al. [23])		✓	
Weighted Matrix Factorization (Steck [25], Bailey [1], Gantner et al. [6])		✓	✓
Max Margin Matrix Factorization (Srebro et al. [24])	✓		✓
Item-Weighted PCA (ours)	✓	✓	✓

Recommender System Evaluation

Evaluation Metric and Datasets

Recommendation-based evaluation metric:

$$\frac{1}{d}\sum_{i-1}^{d} \mathsf{AUC}\left(XP'_{j}, y_{j}\right) \qquad \text{(Item AUC-ROC)}$$

$$P' = P - I$$

Datasets

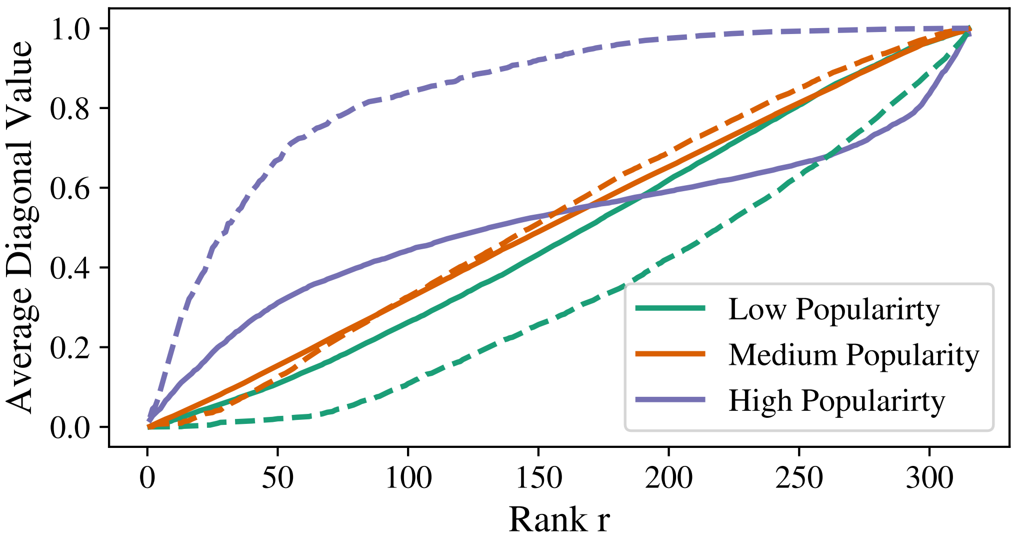
- Last.fm
  - 920 users and 316 artists.
  - Listening counts for each user and artist pair (implicit feedback)
- Movielens
  - 2,000 users and 308 movies
  - Ratings on a 5-star scale (explicit feedback)

Algorithms:

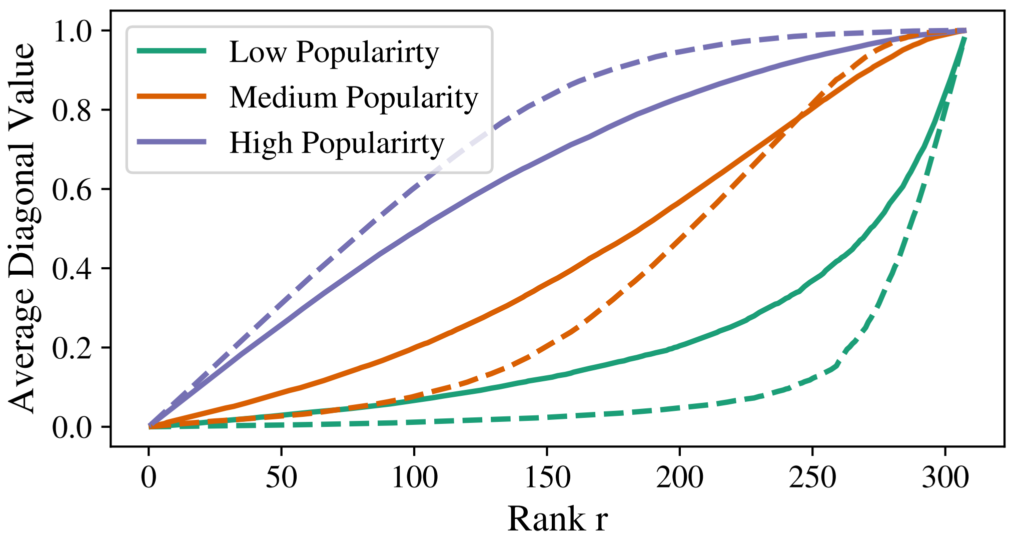
- Item-weighted PCA
- Vanilla PCA
- Column-Normalized PCA

Reduced Specialization

Diagonal value as a heuristic for specialization



LastFM

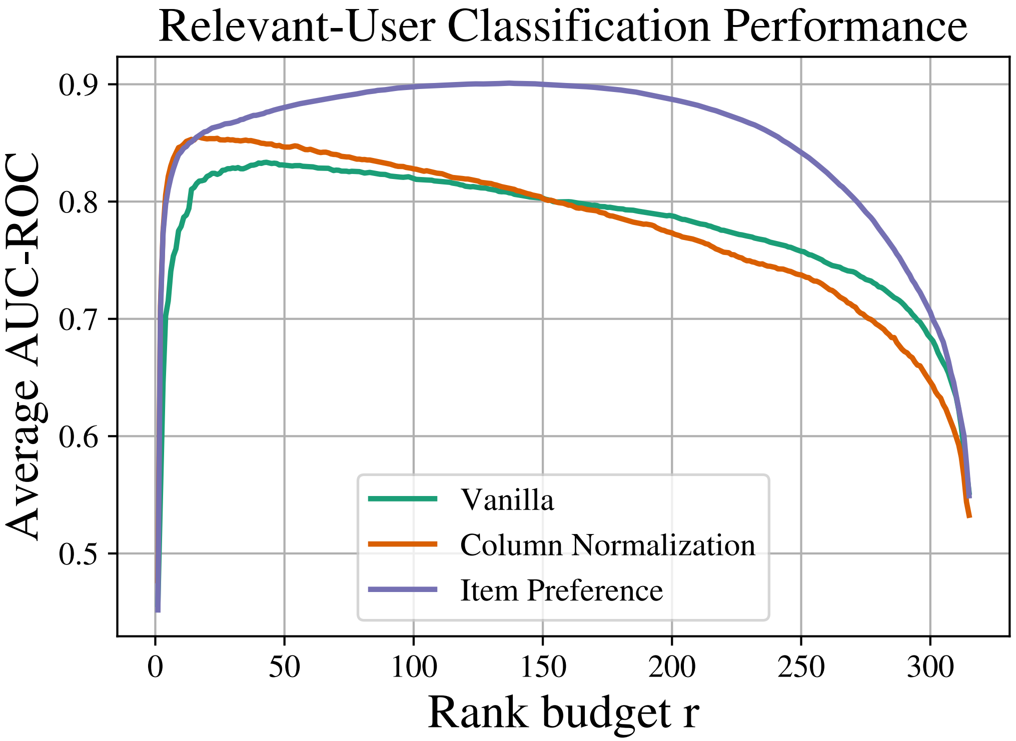


MovieLens

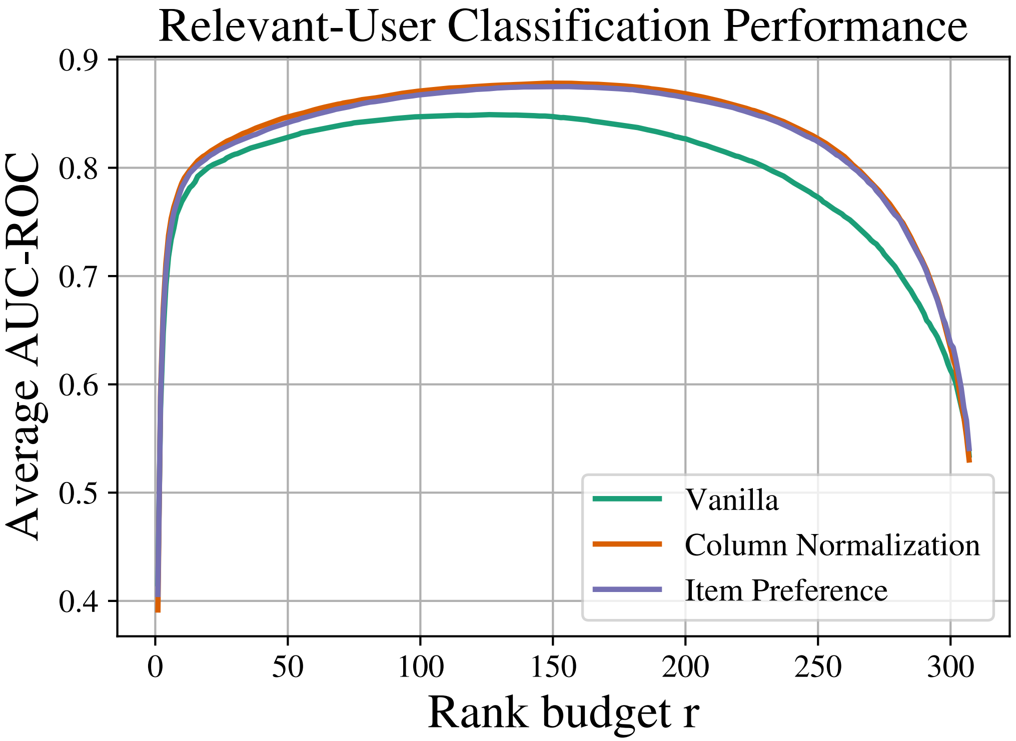
Dashed: Vanilla PCA
Solid: Item-Weighted PCA

Improved User Classification

LastFM

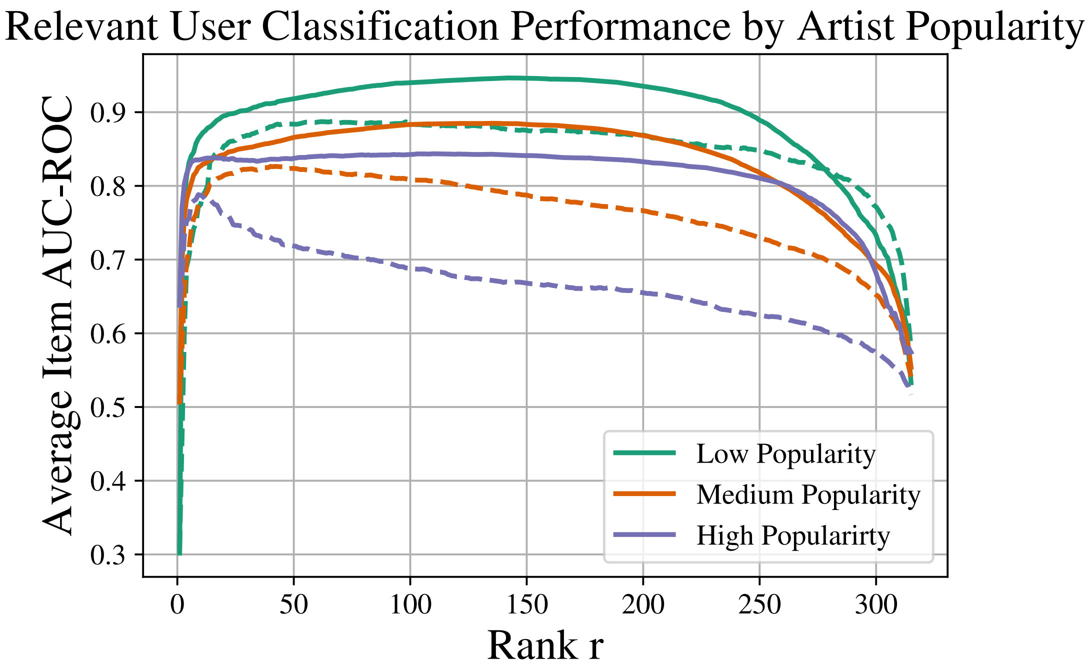


MovieLens

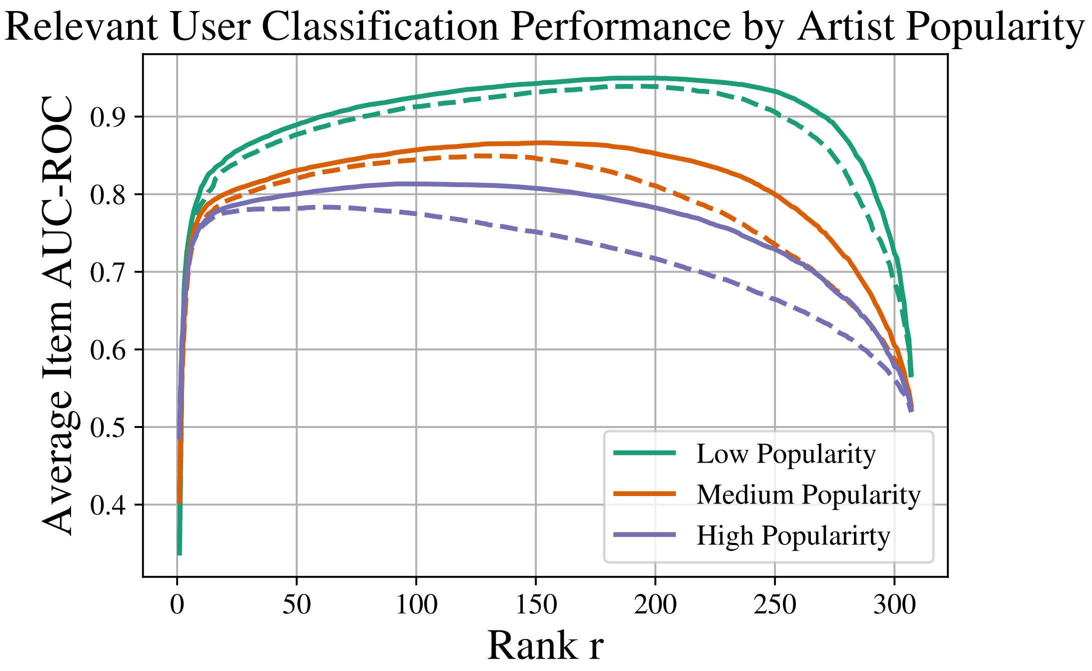


Improvement for Popular and Unpopular Artists

LastFM



MovieLens



Limitations of Item-Weighted PCA

- Solving the SDP runs in O($d^{6.5}$)

- By upweighting unpopular items, Item-Weighted PCA may overfit to noisy data.

- The principal components are not ordered i.e. the solution for rank r+1 is not simply the solution for rank r plus an additional vector.

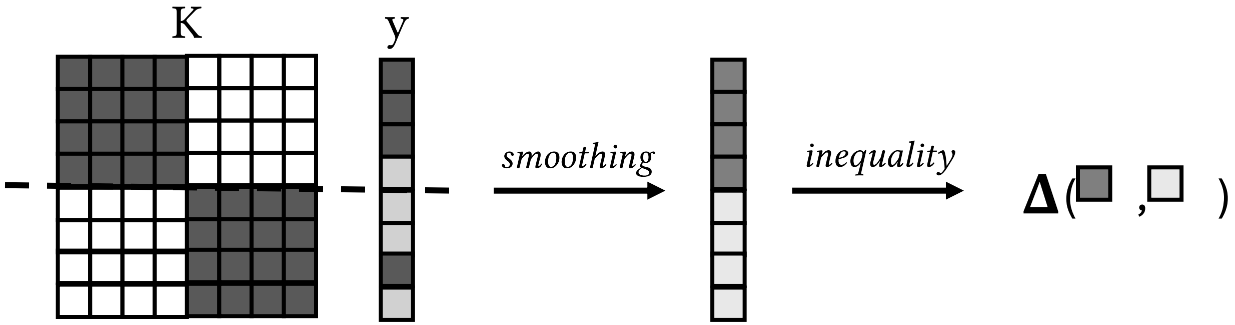
Unfairness in PCA: Takeaways and Future Work

- Identify mechanisms of unfairness underlying PCA as opposed to reliance on external fairness constraints.
- Two mechanisms stemming from popularity disparity
  1. Leading components prioritize popular items
  2. Leading components specialize on *individual* items
- Item-Weighted PCA is an efficient, flexible algorithm.
- Item-Weighted PCA improves recommendations for both popular and unpopular items.

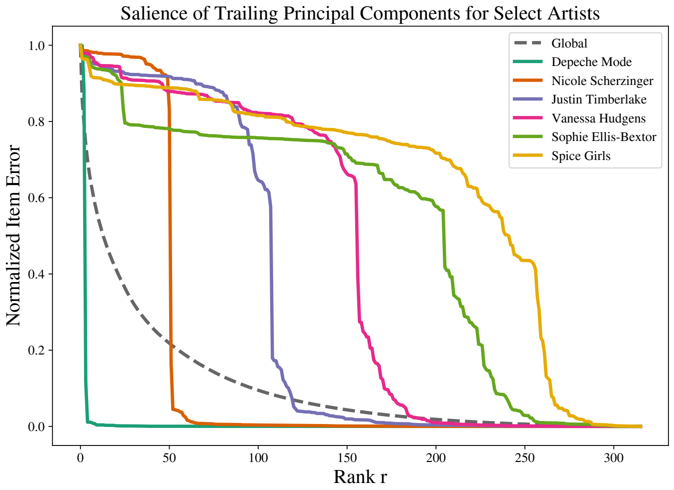
What are additional problems in which we can study underlying unfairness mechanisms instead of using sensitive attribute constraints?

Looking Back

1. Reliance on demographic attributes
   - ➢ Defining group fairness with social networks



2. Don’t help us understand sources of unfairness
   - ➢ Identify mechanisms of unfairness in PCA collaborative filtering



Recommendations for Future Work

- Think about the various appropriate groups for the problem in question

- Understanding fairness in the context of specific models beyond generic definitions.
  - Within representation learning, how can we preserve the richness/idiosyncrasies of each group?

- Deriving mitigation strategies from unfairness mechanisms

Questions?

Group fairness without demographics using social networks
David Liu, Virginie Do, Nicolas Usunier, Maximilian Nickel
FAccT'23
[arXiv 2305.11361]

When Collaborative Filtering is not Collaborative: Unfairness of PCA for Recommendations
David Liu, Jackie Baek, Tina Eliassi-Rad
In Submission
[arXiv 2310.09687]

dliu18.github.io

liu.davi@northeastern.edu

@dayvidliu