TDW on Trustworthy AI

Theme Development Workshops

March 6th, 2026
Paris, France (Hybrid event)

There is No Universal Fairness

Lessons from Link Prediction and Document Classification

Charlotte Laclau
Associate Professor @Telecom Paris, IP Paris

Funded by the European Union

Outline for today

- What is fairness?
- Fair Link Prediction
- Fair Document Classification
- Concluding Remarks

What does it mean to be fair ?

It depends ...

A matter of perspective

Social Sciences

A system is fair if people perceive it as being fair

A matter of perspective

Philosophy

Fairness asks what is morally right, focusing on how decisions should be made.

A matter of perspective

Law

Fairness is about preventing discrimination and limiting *unjustified* unequal outcomes.

A matter of perspective

Quantitative Fields

Fairness is addressed through a pure quantitative perspective

Fairness in Supervised Learning

$$(\underbrace{\text{features}}_{X}, \underbrace{\text{sensitive attribute}}_{A}, \underbrace{\text{label}}_{Y}) \sim \mathbb{P} \quad \text{on } \mathcal{X} \times \mathcal{A} \times \mathcal{Y}$$

Goal: learn a function $h_\theta : \mathcal{X} \to \mathcal{Y}$ such that

$$\min_\theta \ \mathbb{E}_{(X,Y)\sim\mathbb{P}}\big[\ell\big(h_\theta(X), Y\big)\big] \quad \text{s.t.} \quad \text{Fairness}(h_\theta) \leq \varepsilon,$$

where $\text{Fairness}(h_\theta)$ can be,

- $\Delta_{DP} : |\mathbb{P}(h_\theta(X) = 1 \mid A = 0) - \mathbb{P}(h_\theta(X) = 1 \mid A = 1)|$
- $\Delta_{EO} : |\mathbb{P}(h_\theta(X) = 1 \mid A = 0, Y = 1) - \mathbb{P}(h_\theta(X) = 1 \mid A = 1, Y = 1)|$

NLP

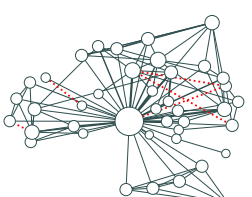
x_i : she is able to assess, diagnose [...]
y_i : nurse
a_i : women


=

→ Surgeon

Graph

x_i : pair of nodes (u, v)
y_i : existence of a link
a_i : ?



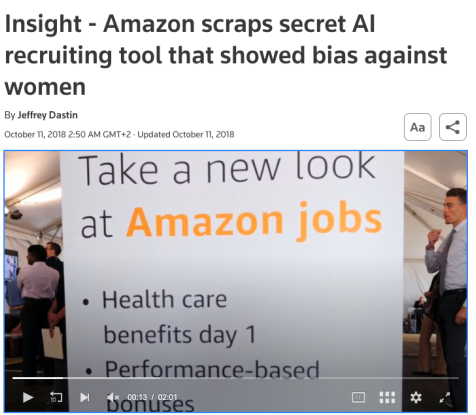
GNN

=

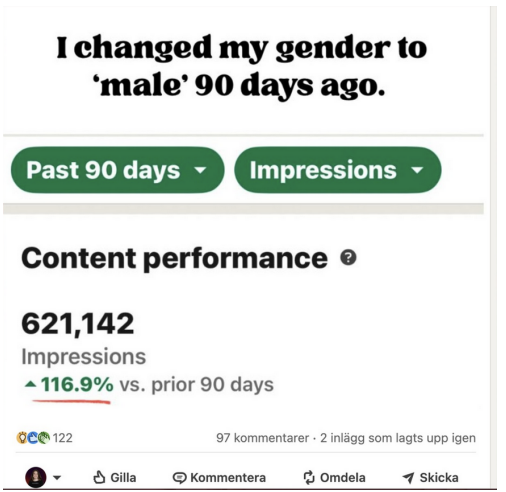
What could go wrong?

What could go wrong?

NLP



Graph



Outline for today

- What is fairness?
- **Fair Link Prediction**
- Fair Document Classification
- Concluding Remarks

Fair Link Prediction[1]

$$g = (\mathcal{V}, \mathcal{E}, F, A), \text{ where } \mathcal{V} = \{v_i\}_{i=1}^{m},\ \mathcal{E} \subseteq \mathcal{V} \times \mathcal{V},\ \mathbf{F} \in \mathbb{R}^{m \times d} \text{ and } A \in \mathbb{R}^{m}$$

Goal: learn a fair link prediction model $h_\theta : \mathcal{V} \times \mathcal{V} \to [0, 1]$.

- Differentiate intra vs. cross group edges.

[1] Laclau C. et al. All of the Fairness for Edge Prediction with Optimal Transport. AIStats 2021

Fair Link Prediction[1]

$$g = (\mathcal{V}, \mathcal{E}, F, A), \text{ where } \mathcal{V} = \{v_i\}_{i=1}^{m},\ \mathcal{E} \subseteq \mathcal{V} \times \mathcal{V},\ \mathbf{F} \in \mathbb{R}^{m \times d} \text{ and } A \in \mathbb{R}^{m}$$

Goal: learn a fair link prediction model $h_\theta : \mathcal{V} \times \mathcal{V} \to [0, 1]$.

- Differentiate intra vs. cross group edges.

Dyadic Fairness (Δ_{DP})

We can define DP for link prediction as

$$\Delta_{DP} : |\mathbb{P}(h(V, V') = 1 | A \oplus A' = 1) - \mathbb{P}(h(V, V') = 1 | A \oplus A' = 0)|$$

[1] Laclau C. et al. All of the Fairness for Edge Prediction with Optimal Transport. AIStats 2021

Connection between homophily and fairness [3]

The homophily mechanism

People generally prefer to befriend others similar to themselves. [2]

[2] [Byrne 1971; McPherson et al. 2001]

[3] Marey L., Perez M., Viard T., Laclau C., TopoFair: Linking Topological Bias to Fairness in Link Prediction Benchmarks, arxiv

Connection between homophily and fairness [3]

The homophily mechanism

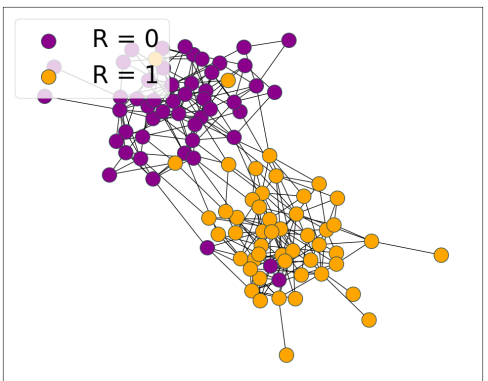
People generally prefer to befriend others similar to themselves. [2]

The homophily mechanism is at the origin of segregation in online networks.


American Sociological Review
1–32
© American Sociological Association 2017
DOI: 10.1177/0003122417705656
journals.sagepub.com/home/asr

Sources of Segregation in Social Networks: A Novel Approach Using Facebook

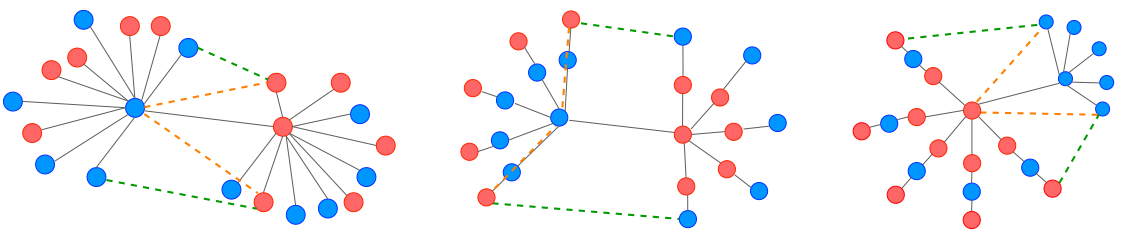
Bas Hofstra,[a] Rense Corten,[a] Frank van Tubergen,[b] and Nicole B. Ellison[c]



[2] [Byrne 1971; McPherson et al. 2001]

[3] Marey L., Perez M., Viard T., Laclau C., TopoFair: Linking Topological Bias to Fairness in Link Prediction Benchmarks, arxiv

Revisiting Dyadic Fairness [4]



- $A = \{$ blue, red$\}$
- We consider two edge predictors ($h^{(1)}$ and $h^{(2)}$), each predicting two edges.
- All graphs have the same assortativity (homophily proxy).
- Both predictors have the same disparate impact.

[4] Marey L., Viard T., Laclau C. k-hop Fairness: Addressing Disparities in Graph Link Prediction Beyond First-Order Neighborhoods, under review.

Defining k-hop Exposure

For $v \in \mathcal{V}$ and $k \in \mathbb{N}^*$, we define the k-hop neighborhood of v as

$$N^{(k)}(v) = \{v' \in \mathcal{V} \mid \sigma(v, v') = k\},$$

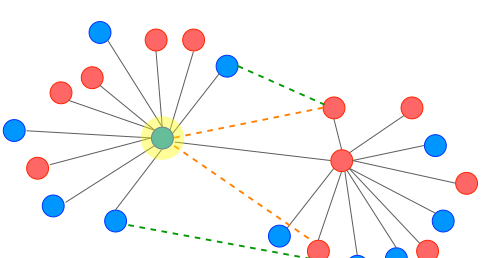
k-hop node exposure

For $s \in \{0, 1\}$, $k \in \mathbb{N}^*$, and $v \in \mathcal{V}$ and $\hat{Y} = h(V, V')$ we define

$$f_s^{(k)}(v) = \begin{cases} \mathbb{E}\left[\hat{Y}\, \mathbb{1}_{\{S'=s\}} \,\middle|\, V = v,\ \sigma(V, V') = k\right], & \text{if } N^{(k)}(v) \neq \varnothing, \\ 0, & \text{otherwise.} \end{cases}$$

$$f_b^1(v) = 5/11, \qquad f_r^1(v) = 6/11$$

$$f_b^2(v) = 5/11, \qquad f_r^2(v) = 5/11$$



Connection with Δ_{DP}

We can rewrite Δ_{DP} across k (nothing fancy just the law of total probability).

$$\Delta_{\text{DP}} = \Big| \sum_{k>0} \sum_{v \in \mathcal{V}} \omega^{(k)}(v) \Big(\frac{f^{(k)}_{\text{same}}(v)}{\mathbb{P}(S = S')} - \frac{f^{(k)}_{\text{diff}}(v)}{\mathbb{P}(S \neq S')} \Big) \Big|$$

with

- $f^{(k)}_{\text{same}}(v)$ and $f^{(k)}_{\text{diff}}(v)$ the within and cross group exposure;
- $\omega^{(k)}(v) = \mathbb{P}(V = v, \sigma(V, V') = k)$

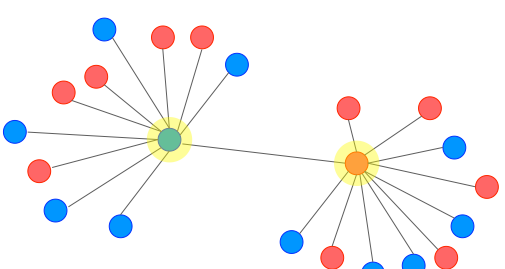
Connection with Δ_{DP}

We can rewrite Δ_{DP} across k (nothing fancy just the law of total probability).

$$\Delta_{\text{DP}} = \Big| \sum_{k>0} \sum_{v \in \mathcal{V}} \omega^{(k)}(v) \Big(\frac{f^{(k)}_{\text{same}}(v)}{\mathbb{P}(S = S')} - \frac{f^{(k)}_{\text{diff}}(v)}{\mathbb{P}(S \neq S')} \Big) \Big|$$

with

- $f^{(k)}_{\text{same}}(v)$ and $f^{(k)}_{\text{diff}}(v)$ the within and cross group exposure;
- $\omega^{(k)}(v) = \mathbb{P}(V = v, \sigma(V, V') = k)$



K-hop Fairness

k-hop fairness.

For a fixed graph distance $k \geq 1$, the fairness gap of a link predictor h is defined as

$$NF^{(k)}(h) = \max_{s \in \mathbb{S}} \max_{s_1, s_2 \in \mathbb{S}} \left| \phi^{(k)}_{s \to s_1}(h) - \phi^{(k)}_{s \to s_2}(h) \right|.$$

In the binary case, $NF^{(k)}(h) = \max_{s \in \{0,1\}} \left| \phi^{(k)}_{s \to 0}(h) - \phi^{(k)}_{s \to 1}(h) \right|$.

1. Look at which group is the most unequally treated.
2. For that group, measure where the access imbalance is the largest.

K-hop Fairness

k-hop fairness.

For a fixed graph distance $k \geq 1$, the fairness gap of a link predictor h is defined as

$$NF^{(k)}(h) = \max_{s \in \mathbb{S}} \max_{s_1, s_2 \in \mathbb{S}} \left| \phi^{(k)}_{s \to s_1}(h) - \phi^{(k)}_{s \to s_2}(h) \right|.$$

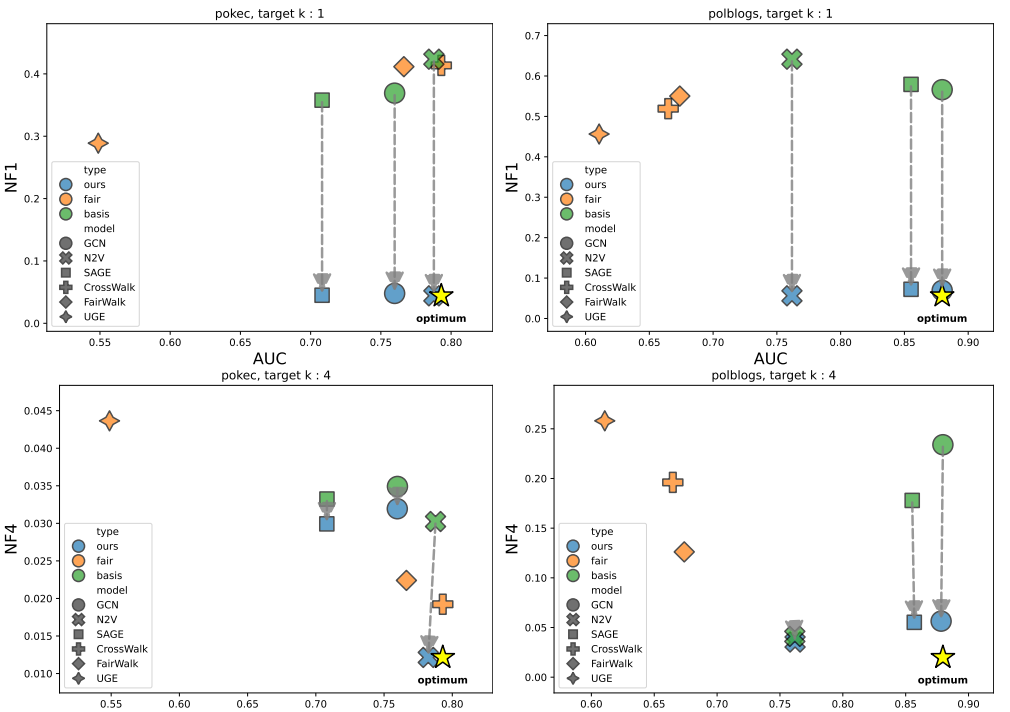
In the binary case, $NF^{(k)}(h) = \max_{s \in \{0,1\}} \left| \phi^{(k)}_{s \to 0}(h) - \phi^{(k)}_{s \to 1}(h) \right|$.

1. Look at which group is the most unequally treated.
2. For that group, measure where the access imbalance is the largest.

Two strategies

rewiring the adjacency matrix or altering the edge probability.

Our findings



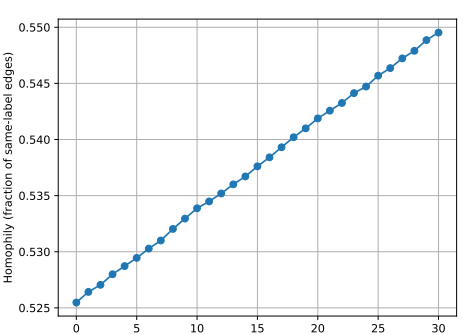
Next problem: Choice vs Induced Homophily

Next problem: Choice vs Induced Homophily

$$\text{Homophily} = \underbrace{\text{Choice}}_{\substack{\text{intrinsic preferences} \\ \text{Who I want to connect to}}} + \underbrace{\text{Induced}}_{\substack{\text{exposure \& structural constraints} \\ \text{Who I am exposed to}}}$$

Dynamic implication

Link prediction shapes the graph and modifies exposure patterns.



Choice vs Induced Homophily a Dynamic Process [5]

Model group-pair interactions as a **multidimensional Hawkes process**.

We express $\mathbb{P}\big(\text{interaction } (i,j) \text{ in } [t, t+dt)\big)$ as

$$\lambda_{ij}(t) = \underbrace{\mu_{ij}}_{\text{choice}} + \underbrace{\sum_{(k,l)} \int_0^t \phi_{(k,l)\to(i,j)}(t-s)\, N_{kl}(ds)}_{\text{induced}}$$

where $\phi_{(k,l)\to(i,j)}(u)$ can encode the score of a link prediction model.

[5] Perez M., Romero R., Lijffijt J., Laclau C., How Predicted Links Influence Network Evolution:Disentangling Choice and Algorithmic Feedback in Dynamic Graphs, under review.

Choice vs Induced Homophily a Dynamic Process [5]

Model group-pair interactions as a **multidimensional Hawkes process**.

We express $\mathbb{P}\big(\text{interaction } (i,j) \text{ in } [t, t+dt)\big)$ as

$$\lambda_{ij}(t) = \underbrace{\mu_{ij}}_{\text{choice}} + \underbrace{\sum_{(k,l)} \int_0^t \phi_{(k,l)\to(i,j)}(t-s)\, N_{kl}(ds)}_{\text{induced}}$$

where $\phi_{(k,l)\to(i,j)}(u)$ can encode the score of a link prediction model.
Given $\lambda_{\mathrm{w}}(t)$ and $\lambda_{\mathrm{c}}(t)$ the aggregated within-group and cross-group intensities:

$$B_{\text{inst}}(t) = \frac{\lambda_{\mathrm{w}}(t)}{\lambda_{\mathrm{w}}(t) + \lambda_{\mathrm{c}}(t)}, \quad B_{\text{inst}}(t) \in [0,1]$$

[5] Perez M., Romero R., Lijffijt J., Laclau C., How Predicted Links Influence Network Evolution:Disentangling Choice and Algorithmic Feedback in Dynamic Graphs, under review.

Long-Term Behavior of Instantaneous Bias

How does $B_{\text{inst}}(t)$ evolve over time?

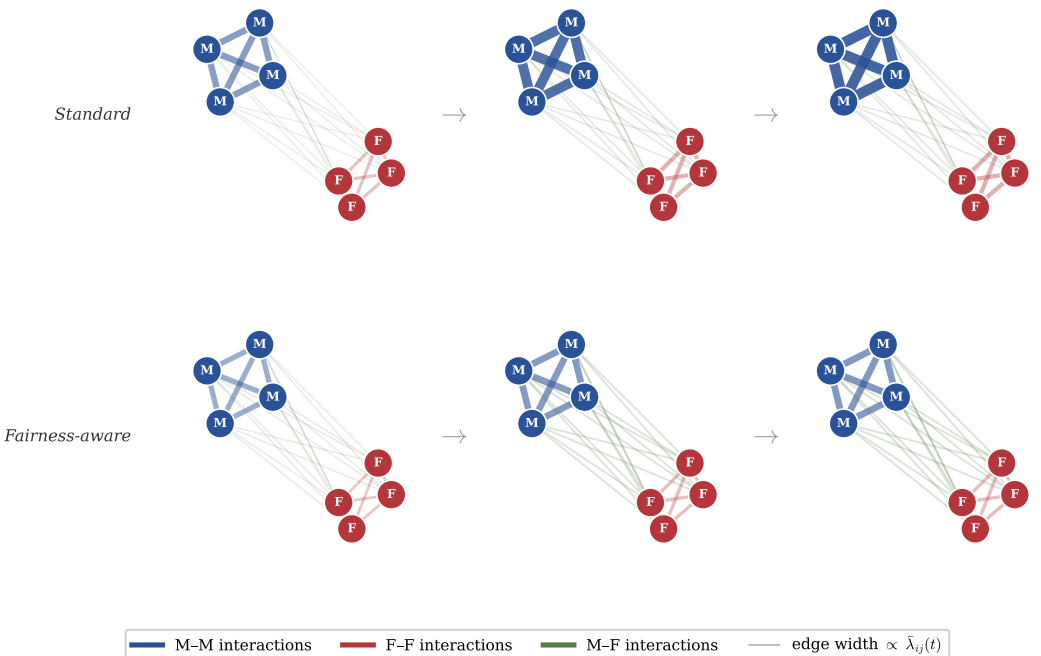
From stochastic dynamics to average evolution

Taking expectations in the Hawkes system yields:

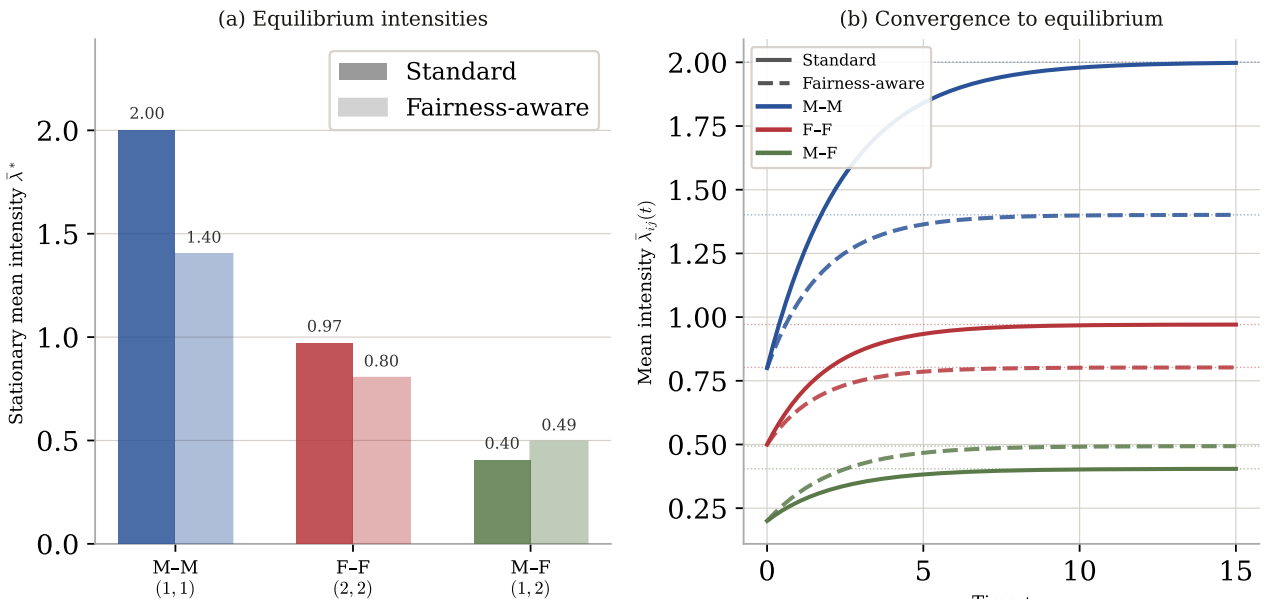
$$\bar{\lambda}_{ij}(t) = \mu_{ij} + \sum_{(k,l)} \int_0^t \phi_{ij,kl}(t-s)\, \bar{\lambda}_{kl}(s)\, ds$$

- The micro-dynamics are stochastic.
- The mean-field system is deterministic.
- Its stability governs the convergence of $B_{\text{inst}}(t)$.

An example: professional network with two groups

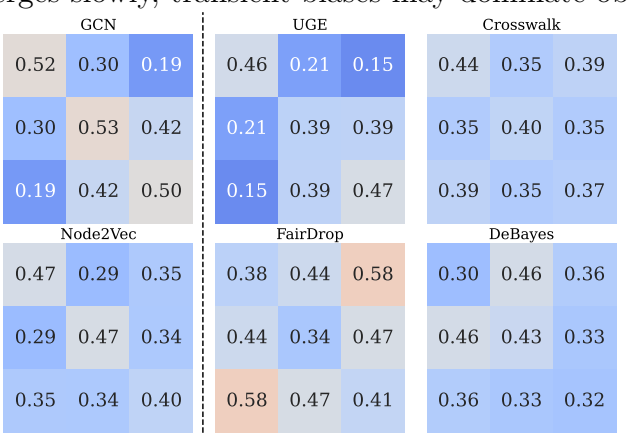


An example: professional network with two groups



Main Theoretical Insights

- **Stability depends on reinforcement strength.**
  If excitation is weaker than decay, interaction intensities converge to a stable regime.
- **Strong reinforcement amplifies asymmetries.**
  When excitation dominates, small initial imbalances can grow and persist.
- **Speed of convergence matters.**
  If the system converges slowly, transient biases may dominate observed behavior.



Outline for today

- What is fairness?
- Fair Link Prediction
- **Fair Document Classification**
- Concluding Remarks

Bias lives in pre-trained models

- Word embeddings trained on large corpora encode social stereotypes
- Language data reflects social and historical inequalities.

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], **Kai-Wei Chang**[2], **James Zou**[2], **Venkatesh Saligrama**[1,2], **Adam Kalai**[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

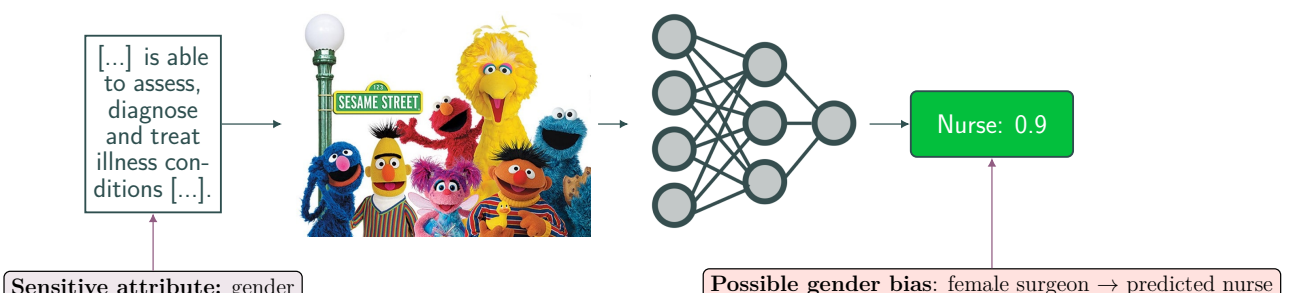
Emily M. Bender[1*], Timnit Gebru[2*],
Angelina McMillan-Major[1], Shmargaret Shmitchell[3]

[1] University of Washington [2] Black in AI [3] The Aether
*These authors contributed equally.

Models reproduce stereotypes carried in the data

Problems

1. Modern NLP pipelines rely on pre-trained LLMs
2. Access to sensitive attribute is not automatic



Fairness Through WDM Minimisation [6]

Wasserstein Dependency Measure

Replace KL by a transport distance between joint and product of marginals:

$$I_{\mathcal{W}} \;=\; \mathcal{W}\Big(p(X,Y),\, p(X)\,p(Y)\Big),$$

where $\mathcal{W}$ is the Wasserstein distance.

Studying the relation betwen I_W and fairness metrics.

[6] Leteno T., Perrot M., Laclau C. et al. Fair Classifier via Transferable Representations, JMLR 2025

Fairness Through WDM Minimisation [6]

Wasserstein Dependency Measure

Replace KL by a transport distance between joint and product of marginals:

$$I_{\mathcal{W}} = \mathcal{W}\Big(p(X,Y),\, p(X)\,p(Y)\Big),$$

where $\mathcal{W}$ is the Wasserstein distance.

Studying the relation betwen I_W and fairness metrics.

We have shown that:

- ✓ $I_W(\hat{Y}, A)$ is a linear combination of DP
- ✓ $I_W((\hat{Y} = Y)|Y = y, A|Y = y)$ is a linear combination of EO

[6] Leteno T., Perrot M., Laclau C. et al. Fair Classifier via Transferable Representations, JMLR 2025

WDM Regularization with Predicted Sensitive Attribute

Objective

$$\min_{\theta} \; \mathcal{L}\big(Y, \pi_Y(X;\theta)\big) \; + \; \beta \, I_{\mathcal{W}}\big(\hat{Y}, A\big)$$

✗ Requires A at train and test time.

WDM Regularization with Predicted Sensitive Attribute

Objective

$$\min_{\theta} \; \mathcal{L}\big(Y, \pi_Y(X;\theta)\big) \; + \; \beta \, I_{\mathcal{W}}\big(\hat{Y}, A\big)$$

✗ Requires A at train and test time.

Predict $\hat{A}$ from X to remove dependence on true A.

$$\hat{A} = h_A(\mathrm{Enc}(X)) \quad \text{i.e.} \quad X \xrightarrow{Enc} Z \xrightarrow{h_A} \hat{A}$$

Lemma 0.2: Approximation of A
We have that:

$$I_W(\hat{Y}, A) \leq I_W(\hat{Y}, \hat{A}) + 2\sqrt[p]{2}\mathbb{P}(A \neq \hat{A})$$

From Discrete to Continuous Space

Objective

$$\min_{\theta} \; \mathcal{L}\big(Y, \pi_Y(X;\theta)\big) \; + \; \beta \, I_{\mathcal{W}}\big(\hat{Y}, \hat{A}\big)$$

✗ Argmax operator is non differentiable

From Discrete to Continuous Space

Objective

$$\min_{\theta} \; \mathcal{L}\big(Y, \pi_Y(X;\theta)\big) \; + \; \beta \, I_{\mathcal{W}}\big(\hat{Y}, \hat{A}\big)$$

✗ Argmax operator is non differentiable

Ensure independence in the latent space

$$\hat{Y} \to Z_y;\ \hat{A} \to Z_a \quad \text{i.e.} \quad I_{\mathcal{W}}(\hat{Y}, \hat{A}) \to I_{\mathcal{W}}(Z_y, Z_a)$$

From Discrete to Continuous Space

Objective

$$\min_{\theta} \; \mathcal{L}\big(Y, \pi_Y(X;\theta)\big) \; + \; \beta \, I_{\mathcal{W}}\big(\hat{Y}, \hat{A}\big)$$

✗ Argmax operator is non differentiable

 Ensure independence in the latent space

$$\hat{Y} \to Z_y;\ \hat{A} \to Z_a \quad \text{i.e.} \quad I_{\mathcal{W}}(\hat{Y}, \hat{A}) \to I_{\mathcal{W}}(Z_y, Z_a)$$

Are the guarantees holding tight? Let's look at the bound!

From Discrete to Continuous Space

Objective

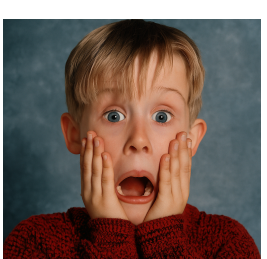
$$\min_{\theta} \; \mathcal{L}\big(Y, \pi_Y(X;\theta)\big) \; + \; \beta \, I_{\mathcal{W}}\big(\hat{Y}, \hat{A}\big)$$

✗ Argmax operator is non differentiable

☼ Ensure independence in the latent space

$$\hat{Y} \to Z_y;\ \hat{A} \to Z_a \quad \text{i.e.} \quad I_{\mathcal{W}}(\hat{Y}, \hat{A}) \to I_{\mathcal{W}}(Z_y, Z_a)$$

Are the guarantees holding tight? Let's look at the bound!



From Discrete to Continuous Space

Objective

$$\min_{\theta} \; \mathcal{L}\big(Y, \pi_Y(X;\theta)\big) \; + \; \beta \, I_{\mathcal{W}}\big(\hat{Y}, \hat{A}\big)$$

✗ Argmax operator is non differentiable

Ensure independence in the latent space

$$\hat{Y} \to Z_y;\ \hat{A} \to Z_a \quad \text{i.e.} \quad I_{\mathcal{W}}(\hat{Y}, \hat{A}) \to I_{\mathcal{W}}(Z_y, Z_a)$$

Are the guarantees holding tight? Let's look at the bound!

Forget the equations: it only depends on two intuitive quantities.

- ✓ **Confidence of predictions**: how clearly one class wins over the others.
- ✓ **Fraction of uncertain cases**: samples fall below the confidence margin.

Next problem: generative models

Next problem: generative models

write two short stories: one about a nurse, the other one about a doctor. For both give them a name.

Next problem: generative models

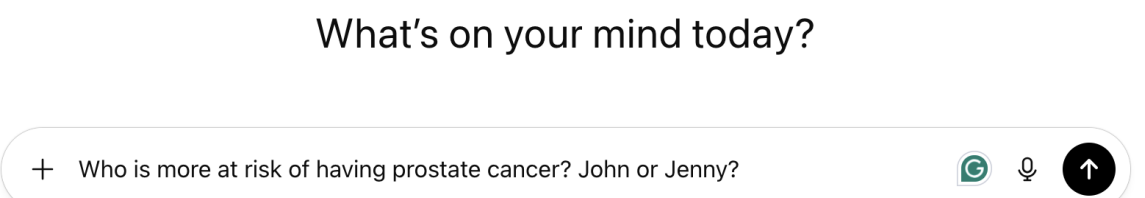
Élise, the Nurse

Élise learned the rhythm of the ward the way others learn music.
She could tell, just by the sound of footsteps, who was anxious and who was healing.
One night, she sat a little longer by an elderly patient who couldn't sleep, listening to stories no chart would ever record.
In the morning, she left quietly, knowing care was often invisible but never small.
That was enough for her.

Dr. Martin, the Doctor

Dr. Martin arrived early, reviewing scans while the hospital was still half-asleep.
He liked the certainty of diagnoses, the moment when confusion narrowed into clarity.
But what stayed with him were not the cases he solved, but the conversations he had to slow down for.
Each decision carried weight, and he felt it deeply, even when he hid it well.
By the end of the day, he reminded himself that medicine was as much about people as it was about answers.

Context is everything



Outline for today

- What is fairness?
- Fair Link Prediction
- Fair Document Classification
- **Concluding Remarks**

Fairness is not a plug-in constraint that can be transferred unchanged across learning settings.

Fairness is not a plug-in constraint that can be transferred unchanged across learning settings.

Reasons

A fairness notion is defined relative to:

- the object being predicted → label, ranking, edges, text etc.
- the intervention point → outputs vs representations vs structure.
- the data-generating system → static vs feedback-driven.

Consequence

The question is not "which metric is best?", but

Which notion matches the system and the harm we aim to prevent?

Any questions ?