Modeling and Mitigating Gender Bias in Matching Problems

A Simulation-Based Approach with Quota Constraints

Florian Wilhelm and Anja Pilz

Why Address Bias in AI Matching Systems?

Anti-discrimination laws (e.g., US Civil Rights Act 1964) and ethical standards prohibit discrimination based on characteristics such as gender, race, religion, or origin.

How can we overcome biases in high-stakes decisions?

Quotas: A Tool to Promote Fairness

- Quotas **enforce minimum participation** of underrepresented groups (e.g., 30% female hires).
- Quotas can **lead to trade-offs with efficiency**, especially when group preferences differ.

How can we mitigate bias without sacrificing efficiency?

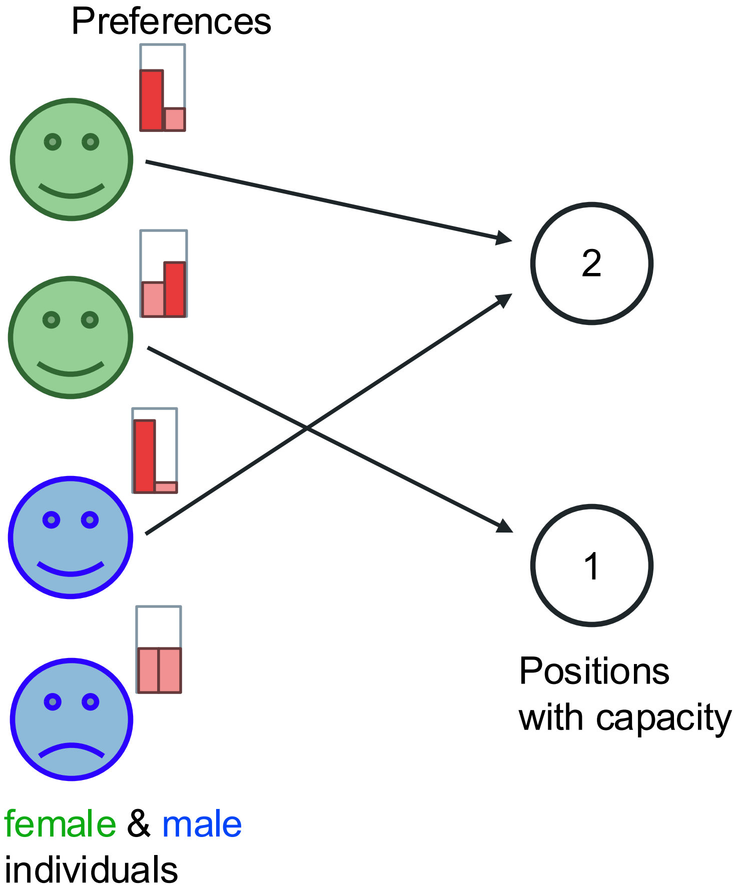
Counterfactual Framework to study Fairness

- *Fairness* is the outcome **under no gender-based bias**.
- But we **can't observe** that unbiased matching directly in real systems.

What happens if we apply quotas under varying gender-specific differences in preferences?

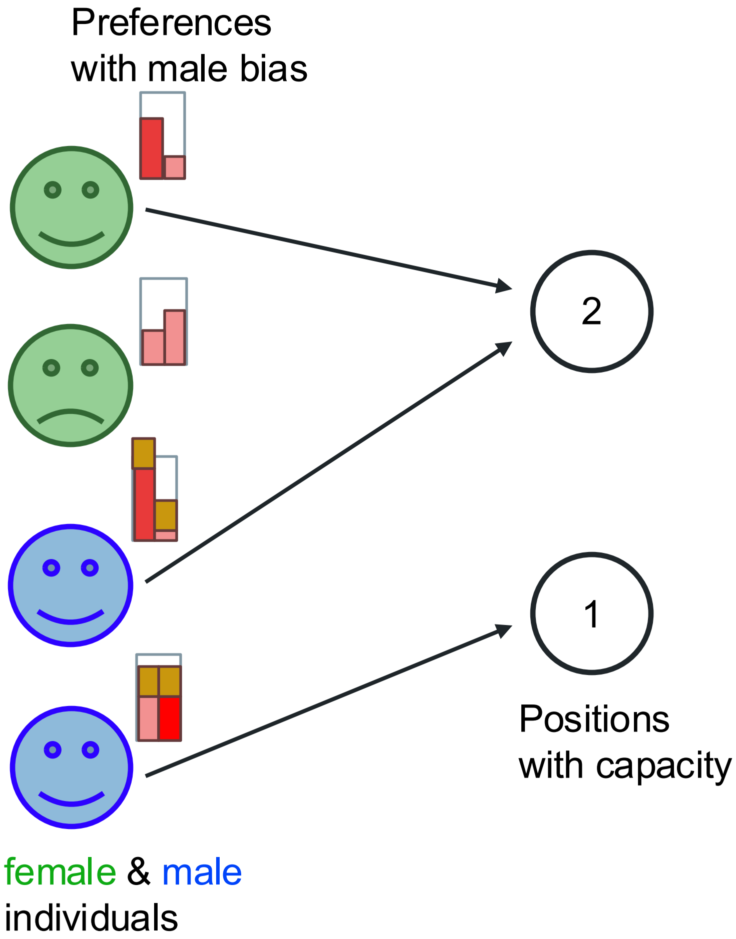
Matching Problem (1/2)

- Positions with certain capacities
- Individuals with gender and preferences for these positions
- Assume *optimal matching* maximises the fulfilled preference based on the matched positions under the Interest-Ability Hypothesis



Matching Problem (2/2)

- A systematic bias β that favours one gender, e.g. male during the matching
- The matching is no longer optimal based on the actual fulfilled preferences
- The *efficiency* η is the ratio of fulfilled preferences in relation to the unbiased case.



Data Generating Process

1. Generate gender $g_i \in \{f, m\}$ for individual $s_i \in S$
$$P(g_i = f) = P(g_i = m) = 0.5$$
2. Sample gender-specific preference priors
$$\boldsymbol{\alpha}^{(g)} \sim \text{Gamma}\left(\alpha_{\text{prefs}}, 1\right)$$
3. Generate individual preferences for $s_i \in S$
$$U_i \sim \text{Dirichlet}(\boldsymbol{\alpha}^{(g_i)})$$
4. Generate capacities for positions $o_j \in O$ with a modified stick-breaking process for even integers.

Gender-specific Differences in Preferences

Total Variation Distance (TVD) to measure differences in the piors of gender preferences

$$\mathrm{TVD}\left(\alpha^{(f)}, \alpha^{(m)}\right) = \frac{1}{2}\sum_{o_j \in O} \left| \frac{\alpha_j^{(f)}}{\left\|\boldsymbol{\alpha}_j^{(f)}\right\|_1} - \frac{\alpha_j^{(m)}}{\left\|\boldsymbol{\alpha}_j^{(m)}\right\|_1} \right|$$

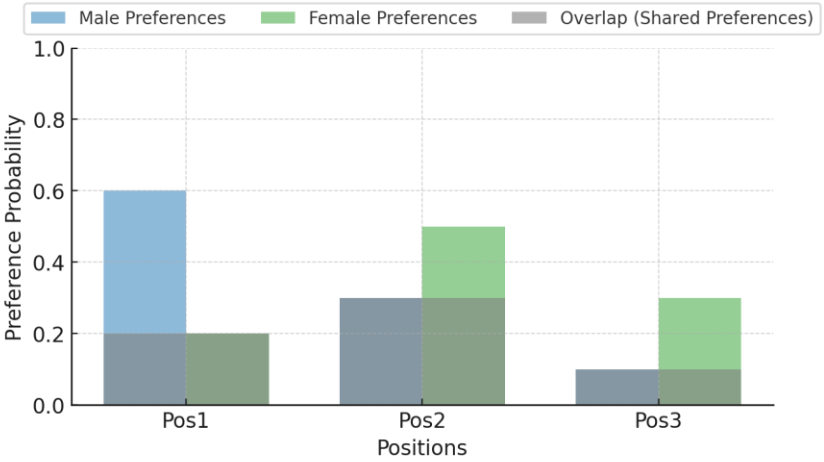
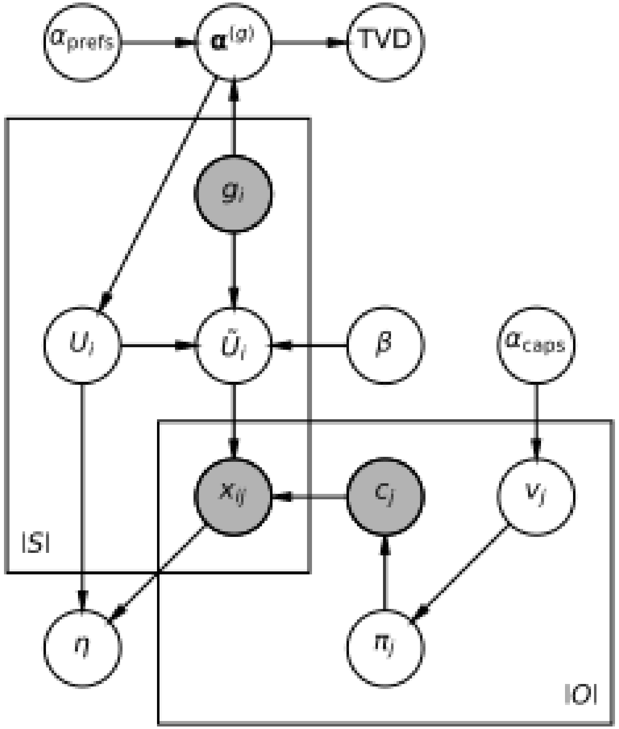


Plate Diagram

For bias β and a quota q perform matching x_{ij} with ILP using the biased preference $\widetilde{U}_i(o_j) = U_i(o_j) + \beta \cdot \delta_m(s_i)$ to study the efficiency η in relation to the TVD of the gender priors $\boldsymbol{\alpha}^{(g)}$



Quotas

- Fixed $q \geq T$ with $T \in \{20\%, \ldots, 50\%\}$ for all positions
- Preference based-quota $\boldsymbol{T}^{(f)}$ based on voting

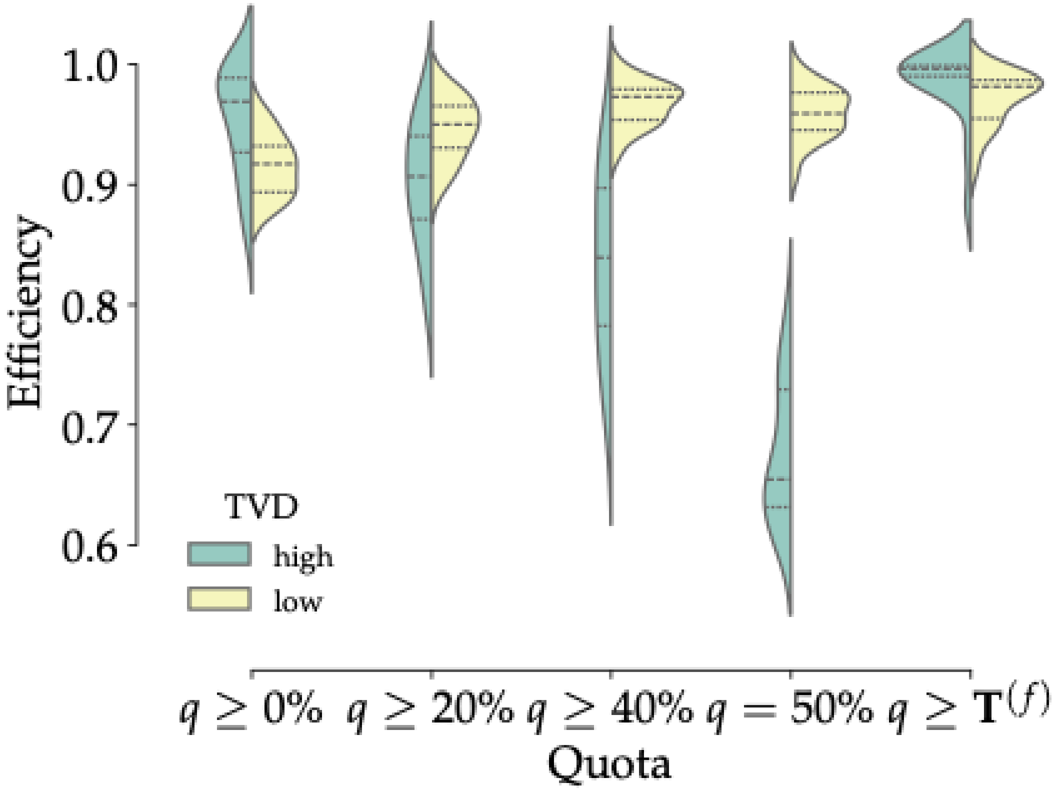
$$\hat{\alpha}_j^{(g)} = \frac{\text{Votes for position } o_j \text{ from gender } g}{\text{Total votes from gender } g}$$

$$T_j^{(g)} = \frac{\hat{\alpha}_j^{(g)}}{\hat{\alpha}_j^{(f)} + \hat{\alpha}_j^{(m)}}$$

Results (1/2)

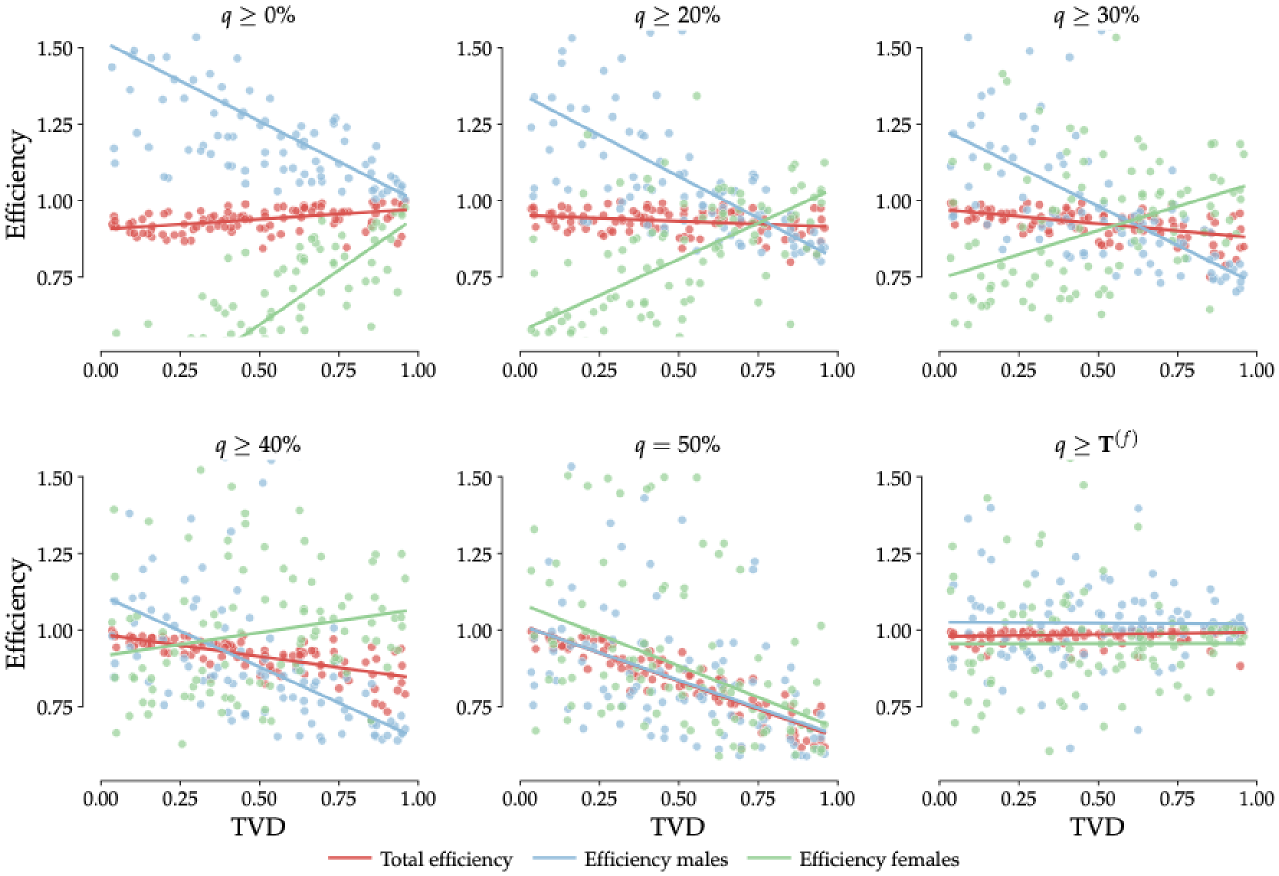
Bias $\beta = 0.3$. For low TVD ≤ 0.2 higher quotas compensate while effiencies decrease for high TVD ≥ 0.8.

Preference-based quotas adjust to varying TVDs.



Results (2/2)

Female, male and total efficiencies for varying TVD and quotas.



Conclusion & Implications

- **Moderate quotas** can enhance both fairness and efficiency when group preferences are similar.
- **Strict quotas** reduce overall efficiency when there is a significant divergence in preferences between groups.
- **Preference-based quotas** are effective in managing high divergence.
- **Our framework** quantifies the trade-offs between fairness and efficiency.
- Code: https://github.com/FlorianWilhelm/gender-bias

Thank you!

Florian Wilhelm
Head of Data Science &
Mathematical Modeling

inovex GmbH

Schanzenstraße 6-20
Kupferhütte 1.13
51063 Cologne

Germany

florian.wilhelm@inovex.de

inovex