Fair Ranking Under Uncertainty

A Major Qualifying Project (MQP) Report
Submitted to the Faculty of
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements
for the Degree of Bachelor of Science in

Computer Science

By:

Marie Tessier
Sai Varun Vadlamudi
Brinda Venkataraman

Project Advisor:

Professor Elke Rudensteiner

With Guidance From:

Kathleen Cachel
Oluseun Olulana

Date: March 2023

Abstract

In today's modern world, digital media and algorithms increasingly control what is seen by users online. This can have serious ramifications when the rankings could perpetuate bias against legally protected groups. These ranking methods must therefore aim for fairness towards such protected groups. However, when necessary demographic information is not available, institutions must look to outside inference algorithms for the information. In our research, we analyze the effectiveness of in-processing and post-processing fair ranking methods under uncertain demographic information. Our results show that using inference algorithms that are less accurate with respect to the protected group with a fair ranking method produces more fair rankings. However, we recommend future researchers to spend more time tuning the parameters of the fair ranking methods.

Executive Summary

Introduction. Digital applications are becoming increasingly relevant in all aspects of our modern world. These applications are often driven by ranking systems, which dominate search engines providing top results for users. The algorithms provide a ranking that can suggest the best library books, top videos to watch, news feed stories, and online store products to buy. These sorts of results may seem innocuous until it is considered that these ranking systems are also used when determining top candidates for job applications, talent searches, loan/mortgage approvals or college admissions.

Background. In order to reduce discrimination against a specified protected group, researchers have explored many methods of fair ranking, including in-processing learning-to-rank models and post-processing fair ranking algorithms. DELTR is an in-processing learning-to-rank model that can produce fair rankings by balancing utility and fairness to reduce the inequality of opportunity and discrimination certain protected groups may face. DetConstSort, a post-processing ranking algorithm, can be tailored to achieve fairness criteria such as equality of opportunity and demographic parity. Both methods of ranking resulted in significant improvements in fairness with respect to the protected groups.

Our Goals. Although it is difficult to find a balance between fairness and utility, it is important to work towards being as fair as possible so that no group is discriminated against. However, there are certain instances when institutions are not able to acquire the information required to produce a fair ranking. In these cases, institutions must turn to inference methods to infer protected attributes. The protected attribute information is necessary, not only for fairness metric calculations, but also when training an LTR model to produce fair rankings with respect to a specified protected group or attribute. Therefore, in our research, we want to understand what happens to the impact on fair ranking methods given inferred protected attributes by calculating the fairness and utility of the rankings produced by various LTR methods.

However, institutions may not always choose to use a fair LTR model to produce fair rankings. Therefore, in our research, we also aim to explore the impact of inferred protected attributes on rankings produced by fair post-processing ranking algorithms. Finally, we want to compare the findings on fair LTR models and on fair post-processing ranking algorithms.

Methods. We first investigate how uncertainty in demographic information affects the performance of both in-processing and post-processing fair ranking methods. To produce inferred demographic information, we chose to use the following three inference algorithms in our ex-

periments: Behind the Name, NameSor, and GenderAPI. Additionally, we use DELTR in our experiments as the in-processing fair ranking method and DetConstSort as the post-processing fair ranking method.

To evaluate the impact on the fairness of rankings produced by DELTR and DetConstSort when given ground-truth and inferred demographic information, we explored three different datasets: all WNBA/NBA players until 2017, Boston Marathon 2017 participants, and Cherry Blossom 2017 race participants. For each dataset, we measured the impact of inference on DELTR and on DetConstSort individually. Then, we compared the results between the two ranking methods when using ground-truth demographic information and when using inferred demographic information. We evaluated all results with four metrics that measure the fairness and utility of rankings.

Conclusions. We found that we produced more fair rankings when using DetConstSort over DELTR. This could be due to the deterministic nature of DetConstSort as a post-processing re-ranking algorithm. DetConstSort allows for fine-tuning the desired characteristics of the ranking outputs, whereas rankings produced by DELTR, a learning-to-rank model, are far less reliable as they are heavily dependent on the methods of training and potential bias present in the dataset.

Additionally, we find that using inference methods that are less accurate with respect to the protected group produced more fair rankings. This conclusion was consistent when using DELTR as well as DetConstSort. However, these results were unexpected, as previous research has come to the opposite conclusions. Due to the time constraints on our research, we were not able to fine tune the gamma parameter of DELTR. Having a poorly fitted model could have had a significant effect on how our datasets were ranked. For future work, experimenting to find the optimal gamma value and number of iterations used when training the models will be critical for each individual dataset used. This will help to create models that are not over or under fitted to the data, allowing for more significant results.

We would also recommend that future work incorporate experiments on other protected attributes such as race, religion, age, etc. Our deliverables include the results and analysis across our three experiments on the WNBA/NBA, Boston Marathon 2017, and Cherry Blossom 2017 datasets. In addition, we developed a modular Python package, FairRank, that will allow future researchers to reproduce our experiments and expand the scope of the research.

Acknowledgements

We would like to thank our advisor, Professor Elke Rudensteiner, for her constant guidance over the past six months. Without her invaluable support, we would not have been able to complete this project as successfully. Additionally, we would like to thank Kathleen Cachel and Oluseun Olulana for their knowledge and time. Their feedback and attention to detail allowed us to progress much faster in the project. In particular, we are grateful to Olu for setting aside time each week outside of our scheduled meetings to answer our questions over the course of this project. We would also like to thank WPI PhD student Isaac Zhou for taking the time to train us to use WPI's Turing Cluster. Without his help, we would not have been able to train as many models as we did for experimentation. Finally, we would like to thank Katherine Crighton for her cheerful support in obtaining the funds required for our project.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Digital applications are becoming increasingly relevant in all aspects of our modern world. These applications are often driven by ranking systems, which dominate search engines providing top results for users. The algorithms provide a ranking that can suggest the best library books, top videos to watch, news feed stories, and online store products to buy. These sorts of results may seem innocuous until it is considered that these ranking systems are also used when determining top candidates for job applications, talent searches, loan/mortgage approvals or college admissions.

Recently, numerous researchers have drawn attention to the ranked lists produced by biased machine learning (ML) models, which can exacerbate discrimination and cause unequal exposure of disadvantaged groups [1, 2, 3, 4]. These sources call attention to the fact that many Learning To Rank (LTR) algorithms base their results solely on utility and disregard fairness[1] for a protected group. The protected group[2] is a demographic of people of which it is illegal to discriminate against. In order to reduce discrimination against a specified protected group, researchers have explored many methods of fair ranking, including in-processing learning-to-rank models and post-processing fair ranking algorithms.

1.2 Related Works

Zehlike et. al. introduce their in-processing learning-to-rank model, DELTR (Disparate Exposure in Learning To Rank) [5]. DELTR can produce fair rankings by balancing utility and fairness to reduce the inequality of opportunity and discrimination certain protected groups may face. In their experiments, they compared the performance of DELTR to pre-processing and post-processing ranking algorithms. Their research allowed them to conclude that DELTR, as an in-processing ranking method, out-performed the pre- and post-processing ranking methods.

Geyik et. al. developed a post-processing ranking algorithm, DetConstSort, that can be tailored to achieve fairness criteria such as equality of opportunity and demographic parity [2]. They claim that DetConstSort and its related algorithms consist of the first large-scale deployed framework for ensuring fairness working heavily in the hiring domain of Linkedin. In order to support their claim, they conducted extensive simulations over different parameter choices, and studied the effect of fairness-aware rankings on both bias and utility measures. This approach resulted in large improvements in fairness metrics while not affecting utility and business metrics.

Additionally, Ghosh et. al. explore how uncertainty impacts the fairness of rankings produced by DetConstSort [1]. In their experiments, they conducted simulations and three case studies to show how inferred demographic information can lead to unfair rankings. Their results allowed them to conclude that using inferred demographic data is not recommended in conjunction with fair ranking algorithms, except in the event that the method of inference was highly accurate.

Most recently, Pietrick et. al. explored the impact of uncertain protected demographic information on rankings produced by DELTR [6]. Their study consisted of experiments completed on the COMPAS dataset, using three gender inference algorithms. Their results suggested that using less accurate inference methods negatively impacted the fairness of rankings produced by DELTR, while more accurate inference methods positively impacted the fairness of such

[1]Fairness is mathematically defined and is measured by exposure of protected groups using different metrics.

[2]Some common protected attributes include race, religion, national origin, gender, marital status, age, and socioeconomic status.

rankings. Their final recommendation was consistent with that of Ghosh et. al; it is imperative to carefully evaluate any chosen inference method to ensure its credibility and accuracy.

1.3 Our Goals

Although it is difficult to find a balance between fairness and utility, it is important to work towards being as fair as possible so that no group is discriminated against. However, there are certain instances when institutions are not able to acquire the information required to produce a fair ranking. For example, the Consumer Financial Protection Bureau states that The Equal Credit Opportunity Act (ECOA) and Regulation B generally prohibit a creditor from inquiring "about the race, color, religion, national origin, or sex of an applicant or any other person in connection with a credit transaction" [7]. In these cases, institutions must turn to inference methods to infer protected attributes. The protected attribute information is necessary, not only for fairness metric calculations, but also when training an LTR model to produce fair rankings with respect to a specified protected group or attribute. Therefore, in our research, we want to understand what happens to the impact on fair ranking methods given inferred protected attributes by calculating the fairness and utility of the rankings produced by various LTR methods.

However, institutions may not always choose to use a fair LTR model to produce fair rankings. As discussed in Section 1.2, previous research has explored fair post-processing ranking algorithms. In our research, we also aim to explore the impact of inferred protected attributes on rankings produced by fair post-processing ranking algorithms. Finally, we want to compare the findings on fair LTR models and on fair post-processing ranking algorithms.

1.4 Our Approach

In this research, we investigate how uncertainty in demographic information affects the performance of both in-processing and post-processing fair ranking methods. To produce inferred demographic information, we chose to use the following three inference algorithms in our experiments: Behind the Name, NameSor, and GenderAPI. Additionally, we use DELTR in our experiments as the in-processing fair ranking method and DetConstSort as the post-processing fair ranking method.

To evaluate the impact on the fairness of rankings produced by DELTR and DetConstSort when given ground-truth and inferred demographic information, we explored three different datasets: all WNBA/NBA players until 2017, Boston Marathon 2017 participants, and Cherry Blossom 2017 race participants. For each dataset, we measured the impact of inference on DELTR and on DetConstSort individually. Then, we compared the results between the two ranking methods when using ground-truth demographic information and when using inferred demographic information. We evaluated all results with four metrics that measure the fairness and utility of rankings.

1.5 Contributions

Our results suggest that inference algorithms with lower accuracy with respect to the protected group produced rankings that were more fair, according to our fairness metrics. As for utility, across all experiments and inference algorithms, there were no discernable patterns that suggested a gain or loss in utility when using inferred demographic information in combination with DELTR or DetConstSort. In future research, we recommend experimenting with the parameters of DELTR and DetConstSort to expand upon the results produced in our research.

The deliverables for this research paper include the results and analysis across our three experiments on the WNBA/NBA, Boston Marathon 2017, and Cherry Blossom 2017 datasets. In

addition, we developed a modular Python package, FairRank, that will allow future researchers to reproduce our experiments and expand the scope of the research.

2 Background

2.1 Related Work

In this section we discuss the related work in the literature to algorithmic fairness, ranking algorithms, and inference methods for legally protected attributes.

2.1.1 Algorithmic Fairness

Zliobaite defines algorithmic fairness as "(1) people that are similar in terms non-protected characteristics should receive similar predictions, and (2) differences in predictions across groups of people can only be as large as justified by non-protected characteristics." [8]. Fairness can be mathematically defined through representation-based and attention-based metrics. Representation-based metrics aim to measure the degree to which subgroups of the population in the ranking are proportionally represented in the top-k rankings and the entire ranking [2]. Studies have shown that users do not pay equal attention to all items in a list [9]. For this reason, we introduce attention-based metrics to implement group fairness [1].

When implementing fairness in conjunction with ranking quality metrics, we can balance fairness with utility [3]. This is more useful than purely fair rankings when attempting to implement algorithmic fairness in a real-world application [2].

2.1.2 Ranking Algorithms

Ranking algorithms use certain criteria to rank items in a dataset. The algorithm relies on learning and estimating parameters based on a specified scoring function. The final ranking produced by LTR will be caused by limited visibility in the position of the protected group where their position is further down in the ranking. This is called position bias and is fairly common in LTR [10]. The problem of position bias can be solved with new algorithms that correct for position by focusing on boosting a protected group to the top of the ranking.

LTR is supervised machine learning, using classification methods to process and learn how to rank a dataset. These classifications aim to categorize data given prior information [11]. Three main classification methods are pre-processing, in-processing, and post-processing learning modules that reduce bias at different stages of training.

Pre-processing typically consists of preparing high-quality training data, often using an approach of removing protected attributes (i.e., race, gender, religion, etc.). Although this technique may seem to remove bias from the dataset, it does not account for proxy variables [12].

Post-processing methods assume that accurate demographic information is given to the ranking algorithm. This is not always the case in real-world datasets. In some cases, such as with credit lending companies, legal barriers prevent the collection of sensitive demographics which results in inferring this information to fill the discrepancy. Inferring the demographic information then causes the post-processing methods to be trained on possibly incorrect or biased data resulting in unfair results [1].

In-processing may be the best method when creating a fair learning to rank algorithm by learning how to control bias. DELTR is an in-processing LTR method that accounts for fairness of both protected and non-protected groups, therefore supporting inferred demographics [**?**]. The DELTR method also uses values that used to tune how much the algorithm takes into account fairness and utility when training a model. The model that is trained is then used to produce a fair ranking of a dataset.

2.1.3 Inference Methods

Demographic Inference Methods, or DIMs, are used to predict certain demographics such as race or gender of individuals based on other given attributes such as location or name of the individual. DIMs are often paired with Learning to Rank algorithms when certain protected attributes are missing. This is important for when institutions need to infer certain information for consumers because they are not allowed by policy to have it directly. For example, as mentioned in the introduction, CPFB is charged with ensuring that lenders are following fair lending laws and addressing discrimination. As a result, "auto lenders and other non-mortgage lenders are generally not allowed to collect consumers' demographic information". However, lenders may still need access to that information and therefore use proxy information to fill in information about consumers' demographic [13]. This proxy information is typically generated by DIMs and as all predicative models and algorithms, they are not right 100% of the time. Specifically In our research we are focusing on DIMs that infer gender based on other attributes such as name.

2.2 Metrics and Algorithms

In this section, we introduce the notation, metrics, and algorithms that are relevant to our research.

2.2.1 Notation

The following is a table of variables used to describe metrics and algorithms for the use of our experiments. The variables are defined here to maintain consistency across formulas that were gathered from varying sources.

Notation	Definition
k	position in a ranking (starting at one)
τ	a ranked list of documents
τ^k	top-k documents in ranking τ
j	protected attribute
j_i	subgroup i in j
$p_{b,a}$	proportion of a present in list b
s_k^τ	utility score of the k^{th} element in ranking τ

Table 1: Summary of Notation.

2.2.2 Group Fairness Metrics

Below we detail the metrics chosen for measuring fairness in a ranking τ. The metric equations have all been rewritten with the notation defined in Table 1 for clarity and consistency.

Skew. Given a ranked list τ, a protected attribute j, and a subgroup j_i, the skew for j_i at position k is given as

$$Skew\ (j_i)@k(\tau) = \frac{p_{\tau^k,j_i}}{p_{\tau,j_i}} \tag{1}$$

where p_{τ^k,j_i} is the proportion of members in subgroup j_i in the top-k rankings in τ, and p_{τ,j_i} is the proportion of members in subgroup j_i in the list τ. For group fairness, skew should be

as close to 1 as possible. This would show that the subgroup j_i is proportionally represented in the top-k rankings as it is represented in the the entire ranking τ. In other words, skew measures group fairness with respect to representation in a given ranking τ. [1, 2]

Average Positional Difference For Skew. Given a list of skews σ_i from ranking τ_i and another list of skews σ_j from ranking τ_j where both rankings are of equal length, the average positional difference for skew of a particular group between two rankings is given as

$$AvgPosDiffSkew(\sigma_i, \sigma_j, \tau_i, \tau_j) = \frac{1}{|\tau|} \sum_{k=1}^{|\tau|} Skew\ (\sigma_j)@k(\tau_j) - Skew\ (\sigma_i)@k(\tau_i) \tag{2}$$

where k is the skew at each position in the ranking and σ_i is the list of skews from the control ranking and σ_j is the list of skews from the experimental ranking. A value of greater than one indicates that skew or the representation of that group increased and a value of less than one indicates that the skew or the representation of that group went down. The closer to 0 means the less of an impact that was made on the skews.

NDKL. Given a ranked list τ, we can derive another representation-based fairness metric known as the Normalized Discounted Kullback-Leibler Divergence, or NDKL Divergence. This metric is defined as

$$NDKL = \frac{1}{Z} \sum_{k=1}^{|\tau|} \frac{1}{log_2(k+1)} * d_{KL}(D_{\tau^k}|D_\tau) \tag{3}$$

where $d_{KL}(D_{\tau^k}|D_\tau)$ is the KL divergence score of D_{τ^k} with respect to D_τ, defined as

$$d_{KL}(D_{\tau^k}|D_\tau) = \sum_j D_{\tau^k}(j) * log_2(\frac{D_{\tau^k}(j)}{D_\tau(j)}) \tag{4}$$

and where Z is defined as

$$Z = \sum_{k=1}^{|\tau|} \frac{1}{log_2(k+1)} \tag{5}$$

For group fairness, NDKL values should be close to zero. Similar to skew, an NDKL value close to zero would indicate that members of the list τ in all subgroups j_i are represented proportionally in the ranking τ^k. [1]

Exposure. Given a ranking τ, we can evaluate the exposure at position k, defined as

$$E_{\tau^k} = \frac{1}{log_2(k+1)} \tag{6}$$

We use exposure to measure the fairness of the ranking τ with respect to attention [3]. We can further define the average exposure per group j_i as follows.

$$E_{\tau^k}(j_i) = \frac{1}{|j_i|} \sum_{|j_i|} E_k \tag{7}$$

Finally, we define the ratio of average exposure between non-protected group j_1 and protected group j_2 in Equation 8.

$$E_{\tau^k}(j_1 : j_2) = \frac{E_k(j_1)}{E_k(j_2)} \tag{8}$$

When $E_{\tau^k}(j_1 : j_2)$ is close to one, we can conclude that the ranking is more fair than when $E_{\tau^k}(j_1 : j_2)$ is closer to zero. When $E_{\tau^k}(j_1 : j_2)$ is greater than one, we can conclude that the non-protected group is over-represented in the ranking. When it is less than one, we conclude that the protected group is over-represented in the ranking.

2.2.3 Ranking Quality Metric

Below we have described the ranking quality metric we have chosen, carried over from previous experiments for this project.

NDCG. Given a ranked list τ, we can compute Normalized Discounted Cumulative Gain, defined as

$$NDCG(\tau) = \frac{1}{Z} \sum_{k=1}^{|\tau|} \frac{s_k^\tau}{log_2(k+1)} \tag{9}$$

where s_k^τ is defined as the utility score of the k^{th} element in τ and Z is defined as in Equation 5 [1]. NDCG is a ranking quality metric that can be used to determine the utility of a given ranking.

Average Positional Difference For NDCG. Given a list of NDCG's σ_i from ranking τ_i and another list of skews σ_j from ranking τ_j where both rankings are of equal length, the average positional difference for skew of a particular group between two rankings is given as

$$AvgPosDiffSkew(\sigma_i, \sigma_j, \tau_i, \tau_j) = \frac{1}{|\tau|} \sum_{k=1}^{|\tau|} Skew\ (\sigma_j)@k(\tau_j) - Skew\ (\sigma_i)@k(\tau_i) \tag{10}$$

where k is the skew at each position in the ranking and σ_i is the list of skews from the control ranking and σ_j is the list of skews from the experimental ranking. A value of greater than one indicates that skew or the representation of that group increased and a value of less than one indicates that the skew or the representation of that group went down. The closer to 0 means the less of an impact that was made on the skews.

2.2.4 DELTR

DELTR (Disparate Exposure in Learning To Rank) is an in-processing method of fair ranking and can produce fairness-unaware and fairness-aware learning-to-rank (LTR) models. All LTR frameworks require an input dataset that includes a numeric score for ranking. DELTR is a list-wise LTR framework created by Zehlike and Castillo that runs on both protected and non-protected groups [5]. This framework ensures equal treatment of members within a specified group using average exposure, while giving the most attention to items at the top of the ranking. By balancing utility and fairness, DELTR can produce a fair ranking that reduces inequality of opportunity and discrimination. To specify the fairness of a model, a user defines a value for the input parameter, gamma. When gamma is equal to zero, the model is "fairness-unaware". When gamma is greater than zero, the model is "fairness-aware".

Fairness-unaware DELTR ranks solely on utility, disregarding fairness for a protected group. This model takes in a dataset with numeric scores that are used for ranking and there is no need for a protected attribute. DELTR uses this input dataset to learn how to produce scores to rank the records in the dataset. A trained model is produced based on the specified numeric score associated with each item in a dataset. This trained model is now capable of producing future scores. This model is then tested with the test dataset, with the numeric score column withheld. The ultimate output is an unordered dataset with predicted scores. This output can

be compared with the known scores that were withheld when initially tested to determine the accuracy. The final ranking that is produced by this model will be focused on utility and does not rank consider equal exposure. A sorting algorithm can then be implemented. See Figure 1.

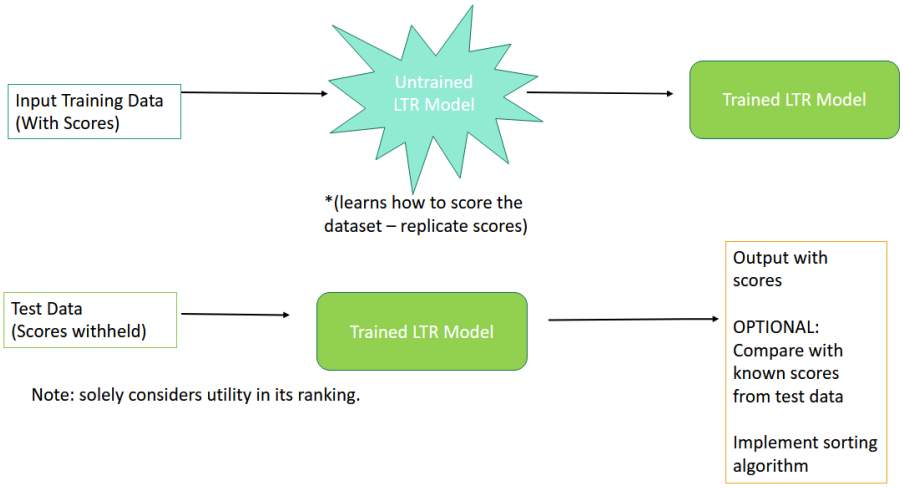


Figure 1: Framework for Vanilla LTR.

Fairness-aware DELTR uses a similar process to the fairness-unaware DELTR model, also known as Vanilla LTR, except instead of basing the ranking only on utility, now this model considers equal exposure of a protected group. Training DELTR with a gamma value greater than one will consider the exposure of the protected value for the top values when ranking a dataset. The model is also more optimized so that the rankings do not compromise the utility of the results [5].

2.2.5 DetConstSort

DetConstSort and its related algorithms were created by Sahin Cem Geyik, Stuart Ambler, Krishnaram Kenthapadi and by the LinkedinIn Corporation as a framework for quantifying and mitigating algorithmic bias in ranking individuals [2]. DetConstSort works as a post-processing re-ranking algorithm. This is to say it takes in an already ranked list by some metric and shuffles the ranking to be more fair. Through this process DetConstSort manages to achive fairness crietria such as equality of opportunity and demographic parity.

DetConstSort works by ranking a list of items based on their scores, while also considering some constraints on how many times each item should appear in the ranking. It works by starting with an empty ranking list and increasing a counter value until at least one item has reached its minimum required count. If multiple items reach their minimum count requirement at the same time, they are ordered based on their scores. Then, the next candidate from each item is inserted into the ranking list, and they are swapped towards earlier positions until they satisfy the constraints. The algorithm repeats this process until the ranking list is full or the constraints cannot be satisfied. In summary, DetConstSort is a sorting algorithm that takes into account some constraints on the appearance of items in the ranking list and tries to maximize the sorting quality while meeting these constraints.

3 Methods

3.1 Objectives

Our goal in our research is to understand what happens to the impact on fair ranking methods, given inferred protected attributes, by calculating the fairness and utility of the rankings produced by in-processing and post-processing ranking methods. We define the following objectives to meet this goal:

1. Research the impact that uncertain, or inferred, demographic information has on the fairness and utility of fair rankings.

2. Research the difference in impact of in-processing ranking algorithms and post-processing fair ranking algorithms.

3.2 Research Questions

To fulfill the research objectives, we research the following set of questions with respect to fairness for in-processing (or learning-to-rank) and post-processing ranking methods:

F1. Given uncertain demographic information, how does the fairness of rankings produced from a fairness-unaware LTR model compare to the fairness of rankings produced from a fairness-aware LTR model?

F2. How does the fairness of rankings produced from a fairness-aware LTR model compare when ranking with ground-truth demographic information and when ranking with inferred demographic information?

F3. How does the fairness of rankings produced by post-processing a ranking from a fairness-unaware LTR differ when using inferred demographics versus ground-truth demographics?

F4. How does the fairness of rankings obtained from a fairness-aware in-processing LTR model, ranked using ground-truth demographics, compare to the fairness of rankings obtained from post-processing a ranking from a fairness-unaware LTR?

F5. How does the fairness of rankings obtained from a fairness-aware in-processing LTR model, ranked using inferred demographics, compare to the fairness of rankings obtained from a post-processing fair ranking algorithm, ranked using inferred demographics?

The next set of questions address the utility of the rankings produced by the same ranking methods:

U1. Given uncertain demographic information, how does the utility of rankings produced from a fairness-unaware LTR model compare to the utility of rankings produced from a fairness-aware LTR model?

U2. How does the utility of rankings produced from a fairness-aware LTR model compare when ranking with ground-truth demographic information and when ranking with inferred demographic information?

U3. How does the utility of rankings produced by post-processing a ranking from a fairness-unaware LTR differ when using inferred demographics versus ground-truth demographics?

U4. How does the utility of rankings obtained from a fairness-aware in-processing LTR model, ranked using ground-truth demographics, compare to the utility of rankings obtained from post-processing a ranking from a fairness-unaware LTR?

U5. How does the utility of rankings obtained from a fairness-aware in-processing LTR model, ranked using inferred demographics, compare to the utility of rankings obtained from a post-processing fair ranking algorithm, ranked using inferred demographics?

3.3 Experimental Design

In this section, we describe how our experiments are designed and the process of selection for the various facets of the design flow.

3.3.1 Ranking Method Selection

We chose to use DELTR as our fair learning-to-rank model in order to continue the research completed by Pietrick et. al. [6]. DetConstSort was chosen because other researchers have previously explored the impact of uncertain demographic information on the fairness of its rankings [1]. Therefore, we were able to build upon previous research to answer our own questions.

3.3.2 Inference Algorithm Selection

To choose an inference algorithm for the purpose of our research, we considered the criteria as defined in Table 2

	Fast	Usability in Code	Credible Source/Has Website	Free (For our purposes)	Accuracy Including Unidentifiable Results Over 80%
Behind the Name	✓	✓	✓	✓	
NameSor	✓	✓	✓	✓	✓
Gender-Guesser	✓	✓		✓	✓
Global-Gender-Predictor		✓		✓	✓
Gender API	✓	✓	✓		✓
NameAPI	✓	✓	✓		

Table 2: Comparison of Inference Algorithms

For each inference method, in addition to ensuring that it could predict gender, it was important to evaluate the various parameters as seen in the table. According to table 2 Behind The Name, NameSor, and GenderAPI were the best choices in terms of fastness, usability, credibility, and price. As for the accuracy we chose to go with one bad, one medium quality, and one good algorithm to see how they compare. These turned out to be Behind The Name, NameSor, and GenderAPI respectively. The accuracies for each algorithm were determined by using them on a datset from data.gov consisting of baby names from Social Security Card Applications [14].

- Fast: If the inference algorithm is able to process atleast 1 name per second.
- Usability in Code: If the inference algorithm has an API key that can easily be obtained and used to send python requests.

- Credibility: If the inference algorithm has a website and has background on how the algorithm returns a gender.
- Free: If the inference algorithm is free to use.
- Accuracy: If the overall accuracy of the inference algorithm is over 80%

For our research purposes, we have chosen three demographic inference methods that predict the race of an individual based on other attributes. The first is Behind the Name. Behind The Name was started by Mike Cambpell in 1996 and is a website used to obtain all aspects of a given name such as the gender, usage (ethnicity), pronunciation, and scripts (spellings of the name in other character sets). In this research, we focus on the gender predictions. Behind the Name operates on a main database of 24,802 names and a submitted name database of 153,316 names that users have submitted and have been approved. This inference method will be used to predict gender based on provided names from chosen datasets.[15]. It can be seen from the table that Behind the Name checks all the criteria except for the fact that it does not have a good overall accuracy. We still decided to use this inference algorithm as our "bad" inference algorithm to see how it compares to the other inference algorithms.

The second demographic inference method is an online API service titled Namesor. Similar to Behind the Name, Namesor also takes in a first and a last name as input and produces the gender and the country of origin/ethnicity of an individual. The Namesor API has been continuously refined over the past 10 years with numerous partnerships in research with universities such as Harvard and Berkeley, scientific groups such as Elsevier, The Lancet, ASME, SSRN, and governmental/international institutions such as ONU, IOM, European Commission. Finally the Namesor AI has processed over 7.5 billion names making it the most accurate name checking technology in the world [16]. NameSor was the medium quality inference algorithm in terms of accuracy.

The final inference algorithm is GenderAPI which is is the biggest platform on the internet to determine gender by a first name, a full name or an email address. GenderAPI achieved the highest accuracy in our chosen inference algorithms and this is because of their advanced multi-layer technology. In their words they are, "not just a simple database lookup. If [they] can't find a name in a specific country, [they] do a global lookup. If [they] can't find a name in a global lookup, [they] perform several normalizations on the name to fix typos and cover all spelling variants [17].

3.3.3 Dataset Selection

When selecting a dataset for this research, there are certain criteria necessary to ensure compatibility with the binary DELTR model. This means that DELTR only recognizes a protected attribute that can be represented by a 1 or a 0. Table 3 below shows the criteria that a dataset must pass to be used for the experiment.

NBA/WNBA Dataset

The first data set we decided to use was a compilation of all NBA and WNBA players. The dataset originally consisted of two seperate datasets, all NBA players, who were male, and all WNBA players, who were female. In each of the data sets we added an additional column titled "Gender" and set them to "M" for male in the NBA dataset and "F" for female in the WNBA dataset. In addition to this the original datasets had a seperate row for each season that a player played that included points scored, PER (player efficiency ration). We had to convert both datasets so that there was only one row per player and in order to do this we condensed the stats across all seasons for a player into the following columns: NumSeasons (number of seasons the player was in), AvgPER (Average player efficiency ratio across all seasons), and

Criteria	Description
File Type (zip/csv)	Ensure compatibility to import with Python.
Who the dataset is describing?	Understand full picture of the data.
How many items are in the dataset?	Ensure there is a sufficient amount of data for training. Greater then 10,000 items.
Are names included in the dataset?	Names are essential for inference methods.
How is the data collected?	Ensure data was collected ethically.
Is there a Ground-Truth Protected Attribute?	Is there a race, gender, etc. attribute accessible.
Is there a numeric scoring feature?	A numerical scoring attribute is essential for training DELTR.
Are there additional numeric scores?	Two numeric scores to come up with additional features.
Is the dataset easy to clean?	Ensure the data can be fit properly for the model.

Table 3: Criteria for choosing a dataset to use within the experiment.

CareerPoints (the total number of points scored by that player across all seasons). After this step we finally merged the WNBA dataset and the NBA dataset into one dataset that was then ready to run our experiment on.

We chose to use this data set because there was a ground-truth protected attribute which was the gender of the player. There was a numeric scoring feature which was the career points scored by the player as well as other scoring features for the LTR model to work with. Finally there was an attribute that we could use to infer gender which was the player name.

Boston Marathon Dataset

The next dataset that was chosen is on participants of the 2017 Boston Marathon Race in Boston, MA. The Boston Marathon is one of the oldest marathons run in the United State, drawing participants from all over the world. Runners abilities range but all have to qualify in order to be a part of the race. The dataset was collected from data scrapped from the official marathon website.

The dataset consists of the name, age, gender, country, city and state (where available), times at 9 different stages of the race, expected time, finish time and pace, overall place, gender place and division place.

As it matched the criteria from Table 3, Boston Marathon was chosen as the second dataset for our experiment. It had ample data points, a gender protected attribute column, and numerous numerical columns that can be used for scoring.

Cherry Blossom Dataset

The final dataset chosen to be used in these experiments consisted of participants in the Cherry Blossom race in 2017. This race is an annual road race in Washington D.C. Some participants ran a five kilometer race, while most ran a 16.09 kilometer (or ten mile) race. The dataset columns include a runner's bib number, their name, sex, age, home city, clock seconds, net seconds, pace seconds, and the event (or distance) that the runner participated in.

We chose this dataset as it had the required types of data and was quite large. This was important as it would allow us flexibility when training models for the experiments. Additionally, it could be more easily compared to the Boston Marathon dataset as they both consist of race data. Finally, it had a significant amount of numerical attributes that were useful for learning scores, and protected attributes of age and gender.

3.3.4 Metric Selection

As discussed in 2.1.1 Algorithmic Fairness, fairness can be mathematically defined through representation-based and attention-based metrics. However, we cannot just evaluate a ranking with respect to fairness; we must balance it with utility for it to be a useful ranking to the user. Therefore, for the experiments in this paper, we choose to evaluate rankings with skew, NDKL,

and the average exposure ratio as our fairness metrics, and NDCG as our utility metric. To find the equations for these metrics, see section 2.2.

3.3.5 Experiment Flow

For each dataset, we produce a distinct set of rankings according to the flow detailed in Figure 2 below.

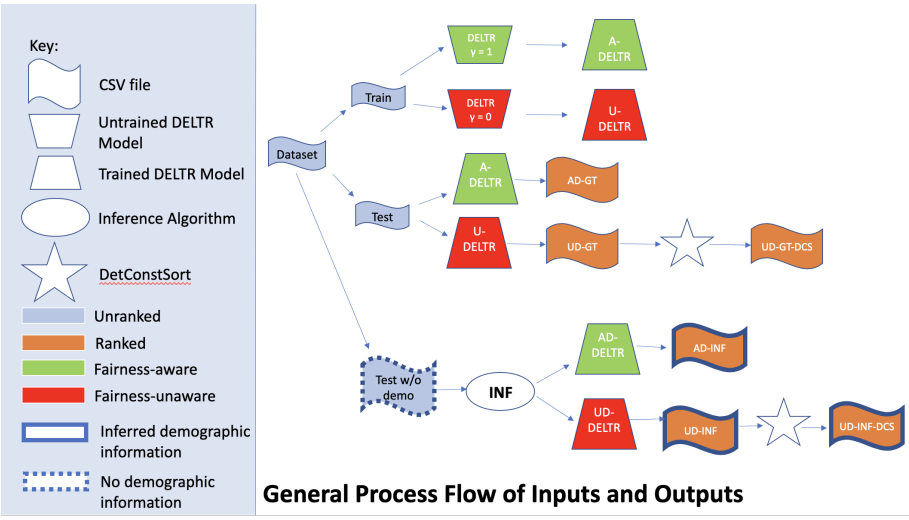


Figure 2: General Process Flow of inputs and outputs to produce distinct rankings for analysis.

We begin by cleaning the dataset so it can be used with DELTR, an in-processing ranking algorithm, and DetConstSort (DCS), a fair re-ranking algorithm. This involves adding columns required for DELTR, rearranging the order of columns, and ensuring the demographic information is represented in a binary format of 0s and 1s. We then split the main dataset into a training split and a testing split. The training split consists of 80% of the original dataset, and the testing split consists of the remaining 20%.

The training split is then used as an input to train two DELTR models. DELTR has a parameter, gamma, that controls how much the demographic information is considered in producing a fair ranking. Of the two models trained in our experiments, one will have gamma = 0 to produce a "fairness-unaware" DELTR model (UD), and the other will have gamma greater than 0 to produce a "fairness-aware" DELTR model (AD). For the experiment, we will initially attempt to train each model on 10,000 epochs.

To produce the first few rankings, we rank the testing split, which has ground-truth (GT) demographic information, with both DELTR models. The rankings produced are named AD-GT and UD-GT. We then pass UD-GT into DetConstSort, to produce a third ranking named UD-GT-DCS.

Inference algorithms get incorporated in the third branch of the process flow diagram. We first duplicate the testing split which is then passed into an inference algorithm. The inference algorithm is used to infer the demographic information, which is then added as a column into the testing split. However, inference algorithms are not able to infer the demographic information of every person in the dataset. Therefore, we produce two versions of the testing split for each inference algorithm: one where the people of unidentifiable gender are given a default gender of male, and one where the default is female. We can refer to these as INF(M) and INF(F), where INF would be an abbreviation of the inference algorithm used. These splits only include inferred demographic information.

The testing splits INF(M) and INF(F) can then be passed as an input to AD and UD. The rankings produced are named AD-INF(M), AD-INF(F), UD-INF(M) and UD-INF(F).

UD-INF(M) and UD-INF(F) are then passed into DetConstSort, to produce rankings named UD-INF(M)-DCS and UD-INF(F)-DCS. This process can be repeated for as many inference algorithms as desired. In these rankings, the ground-truth demographic information is then inserted back in, producing rankings that are now ready for analysis.

The final set of rankings produced for an experiment on a dataset are as follows:

- UD-GT
- AD-GT
- UD-GT-DCS
- UD-INF(M)
- UD-INF(F)
- AD-INF(M)
- AD-INF(F)
- UD-INF(M)-DCS
- UD-INF(F)-DCS

This experimental process can be completed by using the FairRank package; the code and documentation can be found at github.com/svadlamudi2/FairRank.

3.3.6 Analysis Framework

The rankings that were produced at the end of the process flow explained in section 3.3.5 can be compared as in Figure 3. The cells that contain a comparison of the ranking against itself, or duplicate comparisons, are blacked out. Cells that are shaded purple refer to a comparison that can be made to ensure that the rankings were produced properly. Cells that are shaded blue refer to comparisons that are either irrelevant to our research objectives, or that do not make logical sense for comparison.

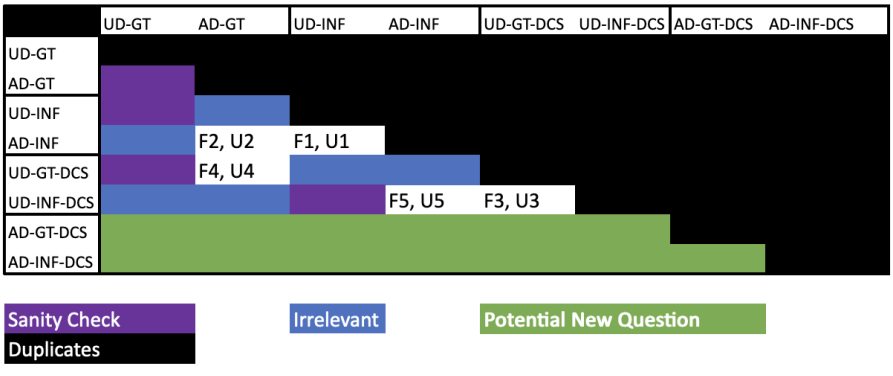


Figure 3: Comparing distinct rankings for research questions.

For example, to answer research questions F1 and U1, we calculate the chosen fairness and utility metrics for rankings UD-INF(M/F) and AD-INF(M/F), then compare.

3.4 FairRank Software

The following section explains how the package we developed for the experiment is structured and how it works. In addition to this, it will explain how the package can be run to produce other experiments. The package can be found at github.com/svadlamudi2/FairRank.

3.4.1 Package Structure

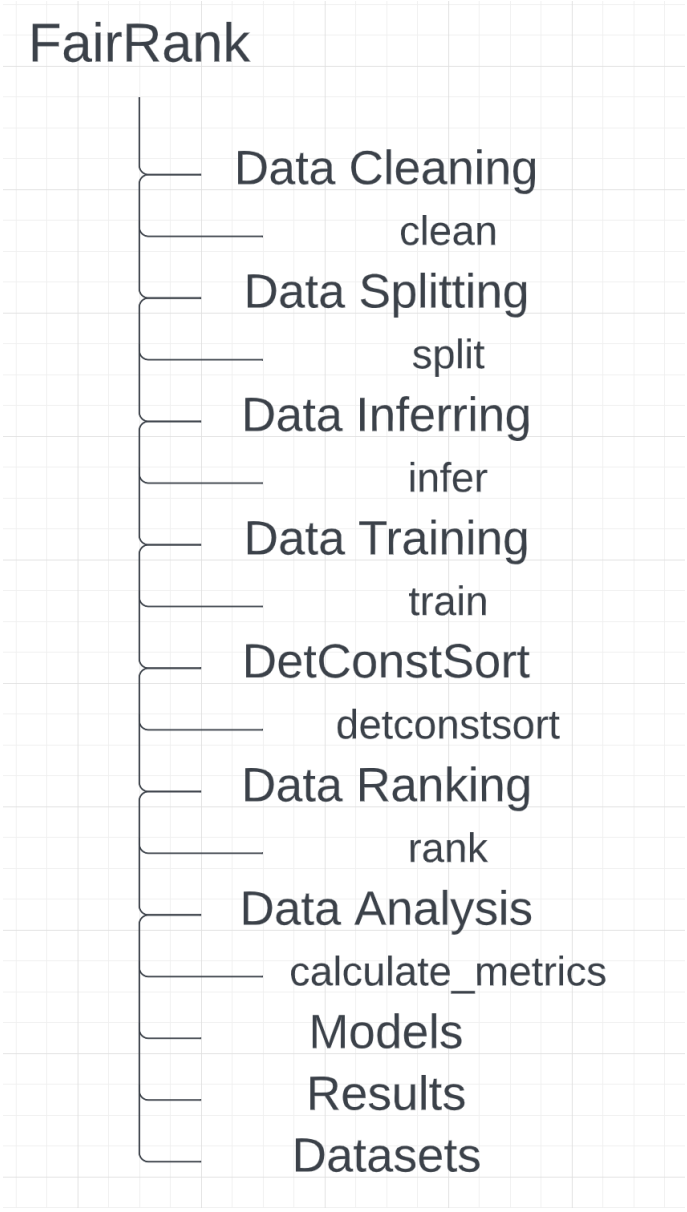


Figure 4: Fair Rank Package Structure

Figure 4 above shows the overall package structure of the Fair Rank software that we have developed in Python to assist in our experiment. It consists of 7 sub packages that have their own python functions.

- Data Cleaning: This sub package helps clean the inputted data and gets it ready so that the LTR model we are using is able to accept it.
- Data Splitting: This sub package helps split the cleaned data into a training and testing split. These split datasets are then saved in the Datasets directory.

- Data Inferring: This sub package helps take in the testing data that was split and uses the inference algorithms described in 3.3.2 to infer gender based on name. The inferred datasets are then written to the Datasets directory.
- Data Training: This sub package helps take in the training data that was split and uses it to train the LTR model that we have chosen as described in 2.2.4. The trained models are then written to the Models directory.
- Data Ranking: This sub package helps take in the testing data, with ground truth demographic information, as well as the inferred datasets, with inferred demographics information, and ranks them using a fairness-aware LTR model and a fairness-unaware LTR model and writes the ranked datasets into the Datasets directory.
- DetConstSort: This sub package utilizes the DetConstSort post-processing ranking algorithm to re-rank the ranked datasets produced by the fairness-unaware LTR model. These re-ranked datasets are then written to the Datasets directory.
- Data Analysis: This sub package takes in all the ranked data sets produced by the "Data Ranking" and "DetConstSort" subpackages and calculates fairness and utility metrics on them. The results are then written to the Results directory.

3.4.2 Running the Code

This section talks about setting up the code to run a full experiment and is broken down into various sections.

Configuring the Settings
The first step is configuring the settings as needed for the experiment. The settings can be found in the settings.json file within the FairRank package. It comprises of 5 main sections: Gender Data Define, Read File Settings, Data Split, Inference Methods, and DELTR options. The following sections will go through each of these sections by setting up the WNBA/NBA experiment.

- **Gender Data Define**

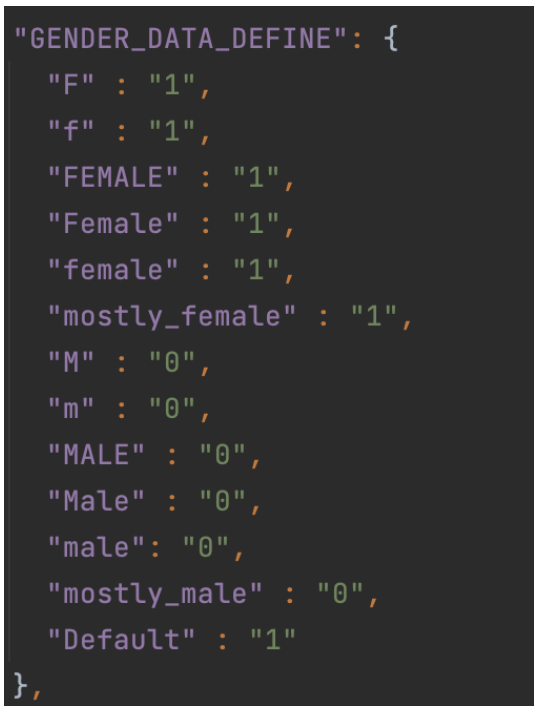


Figure 5: Sample Gender Data Define Settings

Figure 5 shows a sample gender data define segment in a sample experiment. This is used so that the values on the left side that may appear in the data sets of the experiment can be replaced with either a 1 or a 0. This is done for the LTR model that we are using that only accepts 1's and 0's for the protected attribute column. A value of 1 indicates that a group is protected and a value of 0 indicates that a group is non-protected. In the WNBA/NBA experiment, females are the protected group and males are the non-protected group

- **Read File Settings**

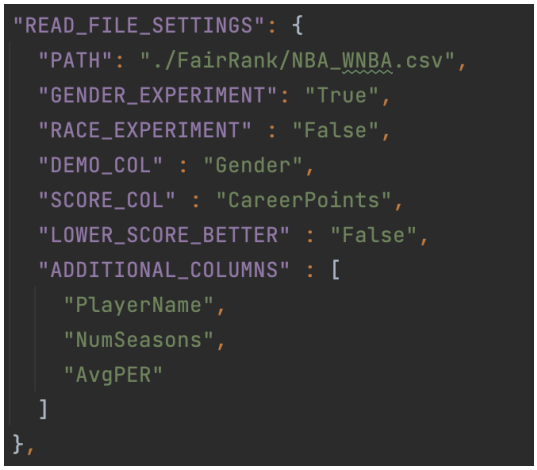


Figure 6: Sample Read File Settings

Figure 6 shows a sample read file settings segment in an experiment. This portion of the settings consists of the path which is the path to the file that you want to run the experiment on. This file should be placed within the FairRank package. Then you must specify whether it is a gender experiment or a race experiment by setting one of them True

and the other to False. The SCORE_COL is is the column name of the scoring feature or what the LTR model is going to learn. In the case of the WNBA/NBA experiment this is the Career Points.

- **Data Split**

```
"DATA_SPLIT" : {
    "TRAIN_PCT" : 0.8
},
```

Figure 7: Sample Data Split File Settings

Figure 7 shows a sample data split setting for the WNBA/NBA experiment. the TRAIN_PCT accepts any values between 0.0 where 0% of the original data will be split to the training data and 1.0 where 100% of the original data will be split to the training data. In the sample code above, 80% of the original data will be split to the training data and 20% will be split to the testing data.

- **Inference Methods**

```
"INFERENCE_METHODS" : {
  "INFER_COL" : "PlayerName",
  "BTN" : {
    "API_KEY" : [],
    "URL" : "https://www.behindthename.com/api/lookup.json?"
  },
  "NMSOR" : {
    "API_KEY" : "",
    "URL" : "https://v2.namsor.com/NamSorAPIv2/api2/json/genderBatch"
  },
  "GAPI" : {
    "API_KEY" : "",
    "URL" : "https://gender-api.com/get?key="
  }
},
```

Figure 8: Inference Methods Settings

Figure 8 shows the different settings used by the inference algorithms in the experiment. INFER_COL is the column name in the original data set that the inference algorithms can use to predict the necessary demographic information. In the case of the WNBA/NBA experiment the column name is "PlayerName". The three different inference algorithms, Behind The Name (BTN), NameSor (NMSOR), and GenderAPI (GAPI) have the same two essential pieces. The API_KEY value is your own individual API_KEY that can be used to make the inference requests to the website. In the case of Behind the Name you need at least two API_KEYS for the code to be functional. The URL, unlike the API_KEYS, should not be touched.

- **DELTR Options**

Figure 9: Sample DELTR Options Settings

Figure 9 shows sample settings for the DELTR options. Gamma can be any value greater than 0.0 where 0.0 is training a fairness-unaware LTR model and any value higher is a fairness-aware LTR model. The "num_iterations" and "standardize" are both values that DELTR requires. SCORE_COLUMN is the column name in the original data set that the LTR model is trying to learn. In the case of the WNBA/NBA experiment the column is titled "CareerPoints". Finally NORMALIZE_SCORE_COLUMN is a value that is either True or False indicating whether or not you want all the values in the scoring column to be normalized to a value between 0 and 1.

4 Results and Analysis

4.1 NBA/WNBA Experiments

Based on the methodology described above, we have ran the experiment on a dataset consisting of WNBA and NBA players. Each player has an associated PlayerName, NumSeasons (number of seasons that they played in their career, AvgPER (their average player efficiency ratio), CareerPoints (the number of points they scored in their career), and Gender.

4.1.1 Cleaning and Splitting WNBA/NBA

The WNBA/NBA dataset was cleaned and ranked according to the score attribute which is the career points. This means that the players with the highest career points will be ranked the highest and the players with the lowest career points will be ranked the lowest. In splitting the data, according to the methodology described above, there was a 80/20 train-test split. This means that 80% of the data was used for training and 20% of the data was used for testing.

4.1.2 Training With WNBA/NBA Data

Using the 80% train split we trained a fairness-aware DELTR model using gamma = 1.0 and a fairness-unaware DELTR model using gamma = 0.0.

4.1.3 Inferring WNBA/NBA Data

According to the methodology described above, Using the 20% test split we developed inferred datasets with inferred genders for the WNBA/NBA players. We used the following inference algorithms:

- Behind The Name
- NameSor
- GenderAPI

When inferring the test data on each inference algorithm there was a male default inferred result and a female default inferred result. This is because, at times the inference algorithms were not able to produce a predicted gender from a name, in these cases we defaulted all the unknowns to male for one data set and all the unknowns to female in another data set.

	Percent Unidentifiable	Accuracy Excluding Unidentifiable Results	Accuracy Including Unidentifiable Results	% Female in Unidentified	% Male in Unidentified	% Accuracy when default is Male	% Accuracy when default is Female
Behind the Name	25.0%	98.0%	73.4%	28.1%	71.9%	91.4%	80.4%
NameSor	15.6%	97.8%	82.3%	17.9%	82.1%	95.3%	85.3%
GenderAPI	2.5%	94.8%	92.5%	41.7%	58.3%	93.9%	93.5%

Table 4: Statistics found when inferring gender using WNBA/NBA Player Names

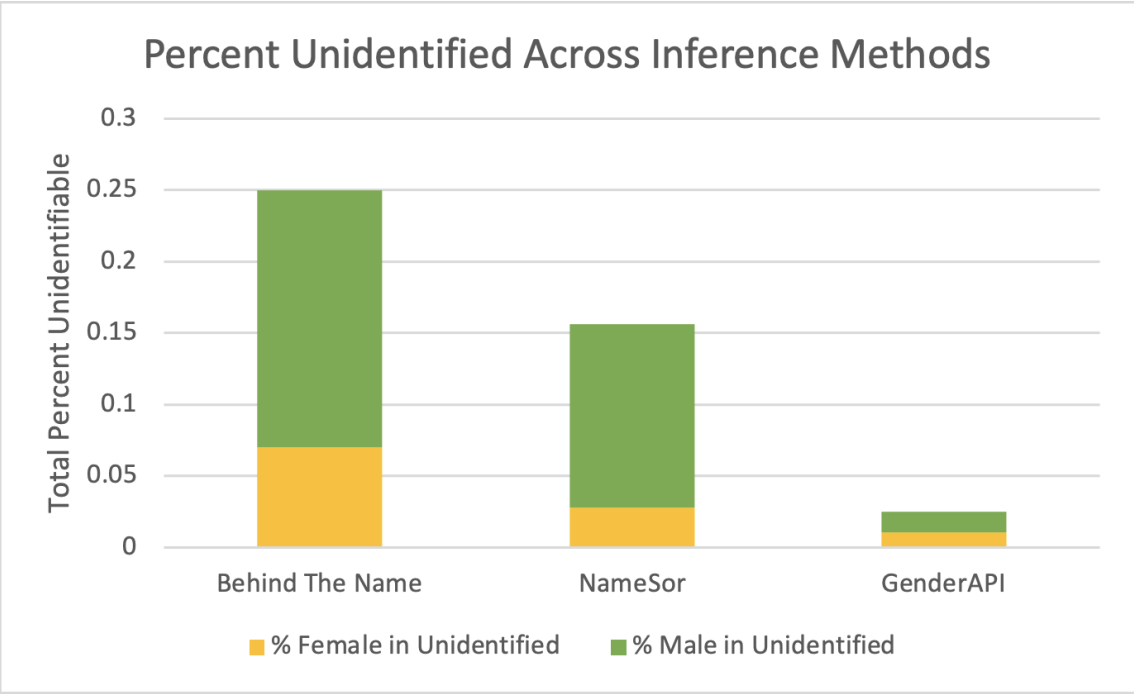


Figure 10: Percentage of unpredictable names WNBA/NBA data set

From the graph in Figure 10 and table 4 we can make the following conclusions about using inference algorithms on the WNBA/NBA data set.

Behind the Name. This inference algorithm was unable to infer the gender of 25.0% of names in the NBA/WNBA dataset. Of those unidentified names, 28.1% were ground-truth female and 71.9% were ground-truth male. When excluding the unidentifiable names, Behind the Name was 98.0% accurate with respect to the ground-truth gender of the names in the dataset. When including the unidentifiable names, Behind the Name was only 73.4% accurate with respect to the ground-truth gender of the names in the dataset. When the unidentified names were assigned male by default, the inference algorithm was 91.4% accurate. When the unidentified names were assigned female by default, the inference algorithm was 80.4% accurate.

NameSor. This inference algorithm was unable to infer the gender of 15.6% of names in the NBA/WNBA dataset. Of those unidentified names, 17.9% were ground-truth female and 82.1% were grounded-truth male. When excluding the unidentifiable names, NameSor was 97.8% accurate with respect to the ground-truth gender of the names in the dataset. When including the unidentifiable names, NameSor was 82.3% accurate with respect to the ground-truth gender of the names in the dataset. When the unidentified names were assigned male by default, the inference algorithm was 95.3% accurate. When the unidentified names were assigned female by default, the inference algorithm was 93.5% accurate.

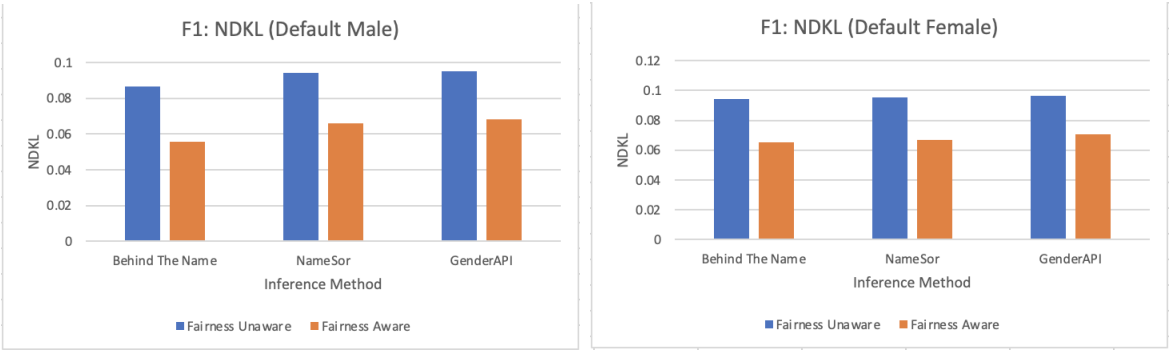
GenderAPI. This inference algorithm was only unable to infer the gender of 2.5% of names in the NBA/WNBA dataset. Of those unidentified names, 41.7% were ground-truth female and 58.3% were ground-truth male. When excluding the unidentifiable names, GenderAPI was 94.8% accurate with respect to the ground-truth gender of the names in the dataset. When including the unidentifiable names, GenderAPI was still 92.5% accurate with respect to the ground-truth gender of the names in the dataset. When the unidentified names were assigned male by default, the inference algorithm was 93.9% accurate. When the unidentified names were assigned female by default, the inference algorithm was 93.5% accurate.

4.1.4 Research Questions

In this section we explore the different fairness and utility research questions described in the methodology with respect to the WNBA/NBA data set.

Fairness Question 1
Given uncertain demographic information, how does the fairness of rankings produced from a fairness-unaware LTR model compare to the fairness of rankings produced from a fairness-aware LTR model?

NDKL



(a) NDKL Male Default (b) NDKL Female Default

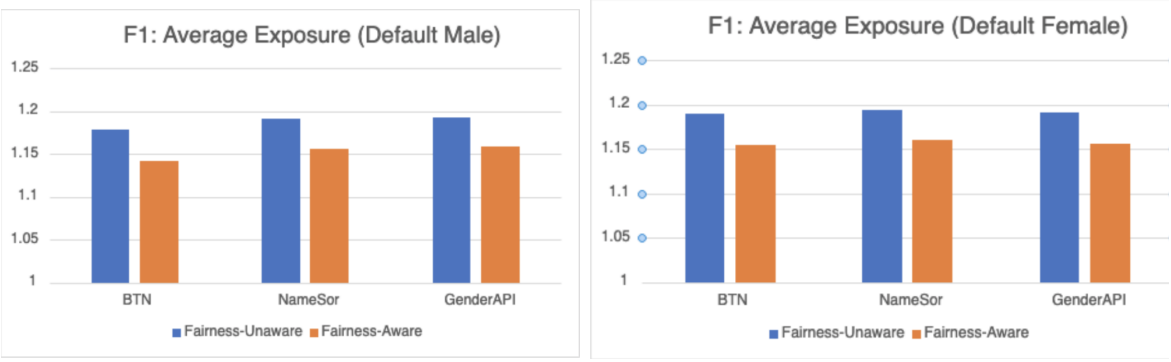
Figure 11: F1, NDKL, WNBA/NBA

From figure 11 it can be seen that for each ranking, the NDKL was lower when ranking with fairness-aware DELTR than when ranking with fairness-unaware DELTR. This means that in each ranking, the subgroups containing males and females were more proportionally represented when DELTR was fairness-aware.

The NDKL was consistently the lowest in rankings where the protected attribute of gender was inferred with Behind the Name. This means that the subgroups containing males and females were more proportionally represented when inferring with Behind the Name.

Across all rankings produced with various inference algorithms, the rankings where unidentified gender was set to male by default consistently had a lower NDKL than when the default was female. For all three inference algorithms, the percent accuracy was lower when assigning unidentified gender to female by default than when the default was male. Therefore, we can conclude that the rankings were more proportionally represented when the default gender for unidentified names was male.

Average Exposure



(a) Avg. Exposure Ratio Male Default (b) Avg. Exposure Female Default

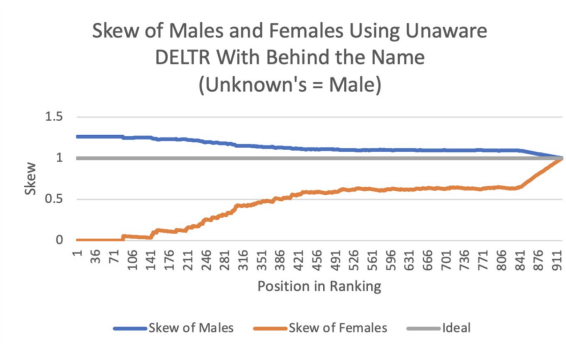
Figure 12: F1, Avg. Exposure Ratio, WNBA/NBA

In figure 12 it can be seen that for each ranking, the Average Exposure was closer to 1 when ranking with fairness-aware DELTR than when ranking with fairness-unaware DELTR. This means that in each ranking, the subgroups were more equally represented when ranking with fairness-aware DELTR.

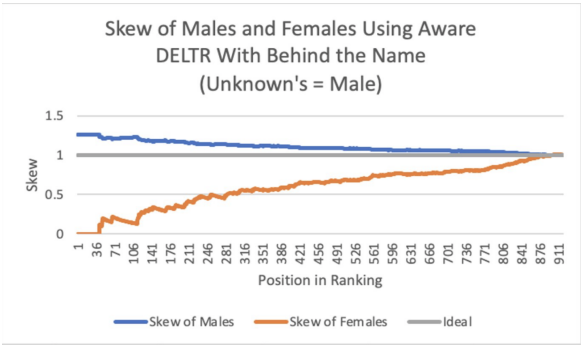
The Average Exposure was consistently closer to 1 in the rankings where the protected attribute of gender were inferred with Behind the Name. This means that the subgroups containing males and females were more equally represented when inferring with Behind the Name.

Across all rankings produced with various inference algorithms, the Average Exposure in rankings where unidentified gender was set to male by default were slightly closer to 1 than the rankings where gender was female by default. This means that the subgroups containing males and females were more equally represented when defaulting to male for unidentified gender.

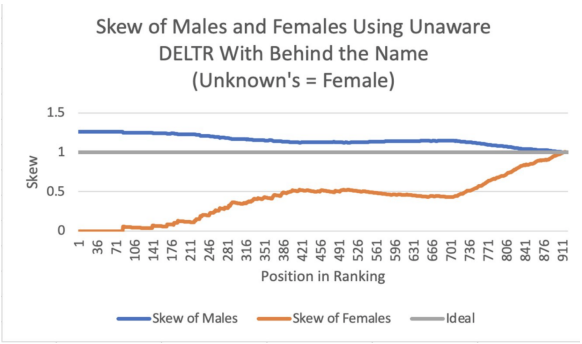
Skew



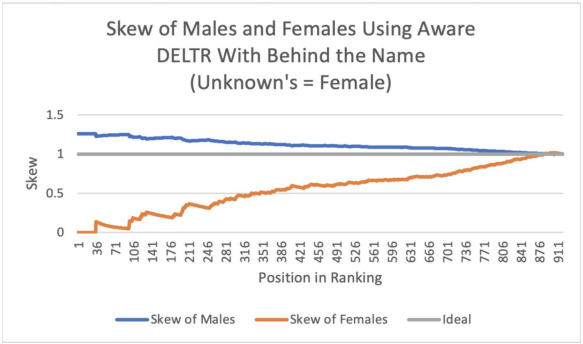
(a) Skew, UD-BTN, Default Male



(b) Skew, AD-BTN, Default Male

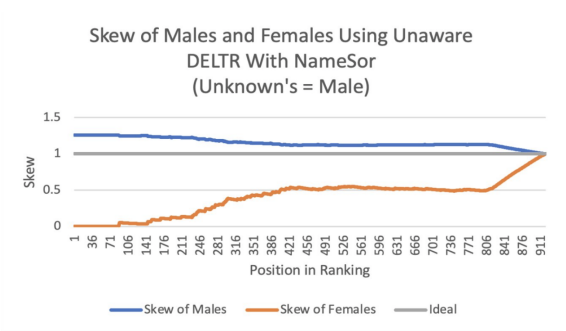


(c) Skew, UD-BTN, Default Female

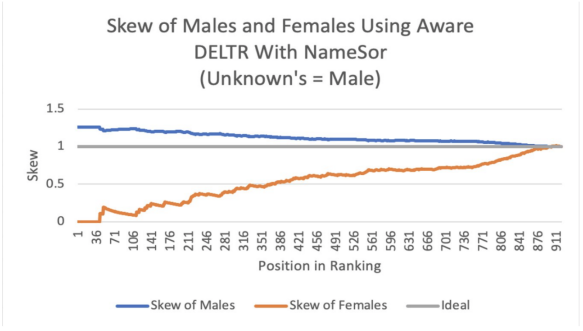


(d) Skew, AD-BTN, Default Female

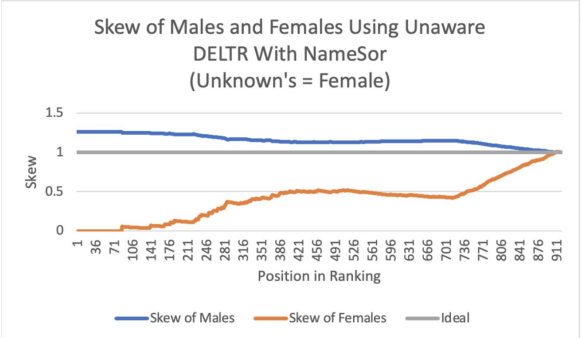
Figure 13: Skew, UD-BTN vs AD-BTN, WNBA/NBA



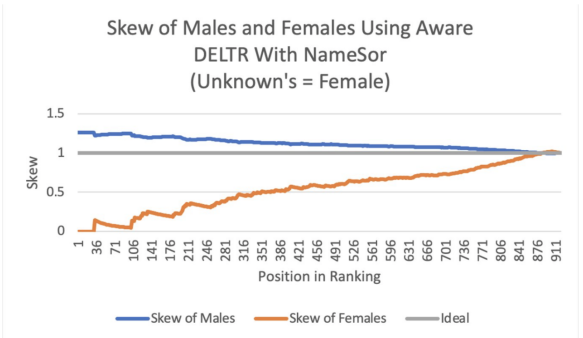
(a) Skew, UD-NSOR, Default Male



(b) Skew, AD-NSOR, Default Male

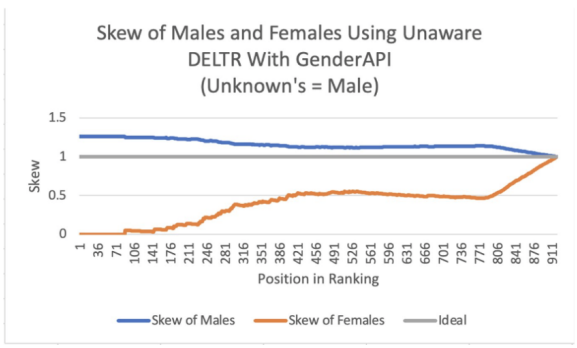


(c) Skew, UD-NSOR, Default Female

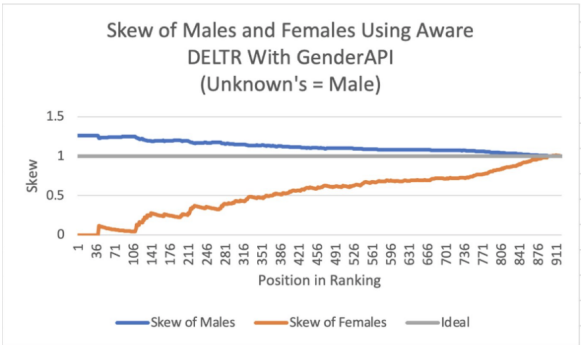


(d) Skew, AD-NSOR, Default Female

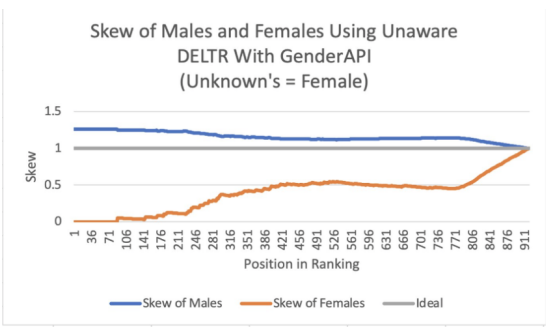
Figure 14: Skew UD-NSOR vs AD-NSOR, WNBA/NBA



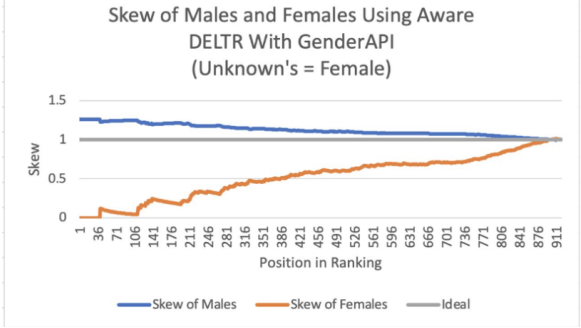
(a) Skew, UD-GAPI, Default Male



(b) Skew, AD-GAPI, Default Male



(c) Skew, UD-GAPI, Default Female



(d) Skew, AD-GAPI, Default Female

Figure 15: Skew, UD-GAPI vs AD-GAPI, WNBA/NBA

In each ranking across all three inference algorithms, the skew increased earlier on in the ranking produced by fairness aware DELTR. Skew is calculated at every position of the ranking for both the male and female groups. A value of greater than one means that a group is over represented at that position in the ranking. On the other hand, a value of less than one means that the group is underrepresented. While the skew is under one for almost the entirety of the rankings for the female group across all inference algorithms, the skew does increase earlier on in the rankings for fairness aware meaning that the the female and male groups are represented earlier on in the rankings using fairness aware DELTR than the rankings using fairness unaware DELTR.

Across all three inference algorithms, the skew for females increases at approximately position 37 in the rankings using fairness-aware DELTR and approximately position 85 in the rankings using fairness-unaware DELTR. This indicates that the representation of both groups remains unaffected across the three inference algorithms of varying quality.

Average Positional Difference in Skew

Table 5 shows the average positional difference of a particular group by comparing two different rankings. The first ranking is produced by the fairness-unaware DELTR and the second is produced by the fairness-aware DELTR. A positive number means that group experienced an increase in the overall skew across all positions meaning that they became more represented. A negative number means that group experienced a decrease in the overall skew across all positions meaning that they became more under represented.

Ranking 1 ⇒ Ranking 2	Males	Females
UD-BTN (Male Default) ⇒ AD-BTN (Male Default)	-0.036855375624285515	0.14179620832290896
UD-BTN (Female Default) ⇒ AD-BTN (Female Default)	-0.035708745215068866	0.13738469869587014
UD-NameSor (Male Default) ⇒ AD-NameSor (Male Default)	-0.035617835190613456	0.13703493433862335
UD-NameSor (Female Default) ⇒ AD-NameSor (Female Default)	-0.035411159429181466	0.13623977654069275
UD-GAPI (Male Default) ⇒ UD-GAPI (Male Default)	-0.03529017931576741	0.1357743214727683
UD-GAPI (Female Default) ⇒ UD-GAPI (Female Default)	-0.03447397159299435	0.13263406965515212

Table 5: F1, Avg. Positional Difference in Skew, WNBA/NBA

According to table 5 across all inference algorithms there was an increase in skew, or representation, for the female group when using a fairness-aware LTR. More specifically, there was a higher average increase in skew for the female group when the unknown demographics from the inference algorithms were defaulted to male.

Conclusion

For the WNBA/NBA experiment, our fairness metrics showed that the rankings were all more fair to the protected group when using a fairness aware DELTR. In addition to this, when using the Behind the Name inference algorithm, the rankings were more fair. More specifically the rankings were more fair across all three fairness metrics when defaulting the unknown values to male.

Utility Question 1
Given uncertain demographic information, how does the utility of rankings produced from a fairness-unaware LTR model compare to the utility of rankings produced from a fairness-aware LTR model?

Table 6 shows the average positional difference in the NDCG of two different rankings. The first ranking is produced by a fairness-unaware LTR model and the second is produced by a fairness-aware LTR model. A positive number means that the NDCG experienced an average increase across all positions meaning that the utility was increased when using a fairness-aware LTR model. A negative number means that the NDCG experienced an average decrease in all positions meaning that the ranking lost utility.

Ranking Comparison	Average Positional Difference
UD-BTN (Male Default) ⇒ AD-BTN (Male Default)	-0.3198512766294554
UD-BTN (Female Default) ⇒ AD-BTN (Female Default)	-0.3198302625840708
UD-NameSor (Male Default) ⇒ AD-NameSor (Male Default)	-0.31986044350890547
UD-NameSor (Female Default) ⇒ AD-NameSor (Female Default)	-0.3198238147118364
UD-GAPI (Male Default) ⇒ AD-GAPI (Male Default)	-0.3198640072279885
UD-GAPI (Female Default) ⇒ AD-GAPI (Female Default)	-0.3198579032036383

Table 6: U1, Avg. Positional Difference in NDCG, WNBA/NBA

From the data in Table 6, it is evident that when implementing a fairness aware LTR model across all inference algorithms, there was a loss in the utility of the ranking. In addition to this, across all inference algorithms, when the unknowns were defaulted to male the NDCG fell by a little bit more meaning that the utility of these rankings was lower than all the others.

Fairness Question 2

How does the fairness of rankings produced from a fairness-aware LTR model compare when ranking with ground-truth demographic information and when ranking with inferred demographic information?

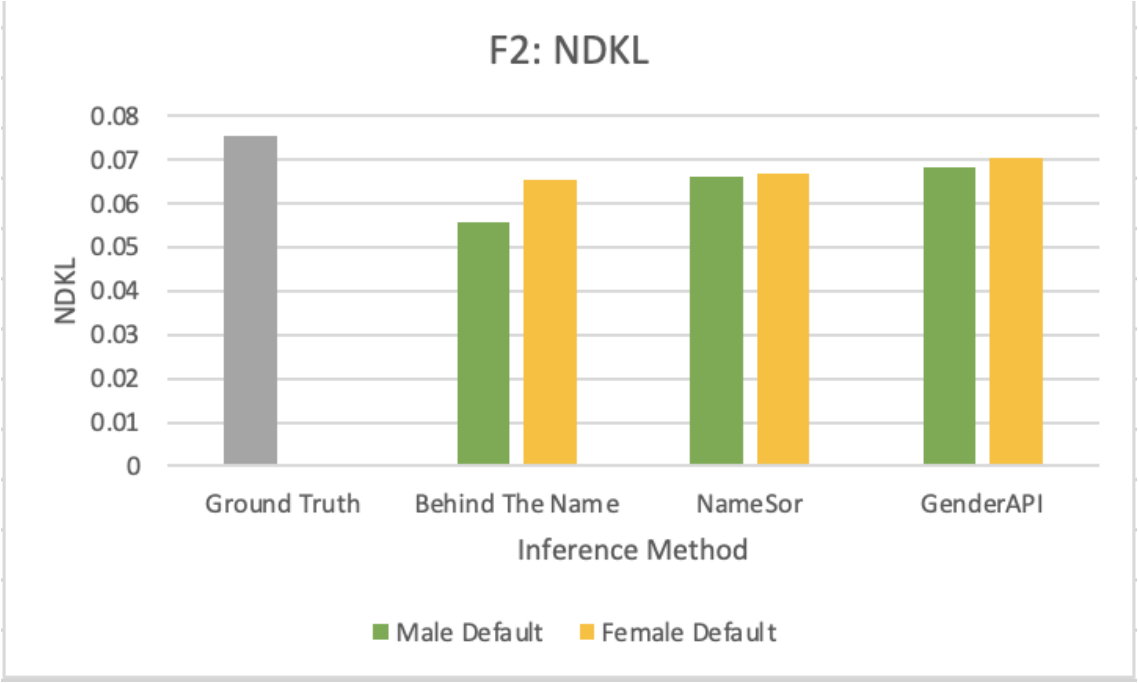


Figure 16: F2, NDKL, WNBA/NBA

From Figure 16 it is clear that the fairness-aware LTR model combined with Behind the Name inference algorithm yields the lowest NDKL meaning the subgroups are the most proportionally represented in this ranking. It is also interesting to note that the fairness-aware LTR model when combined with ground-truth information yielded the highest NDKL meaning the subgroups are least proportionally represented in this ranking.

All three inference algorithms behaved relatively in the same range but Behind the Name performed better than the other inference algorithms having a lower NDKL representing a more fair ranking.

For all three inference algorithms when the unknowns were defaulted to male, this resulted in a lower NDKL. The most drastic difference can be seen for the Behind the Name Inference algorithm that has a 0.1 difference in NDKL when comparing male default to female default.

Average Exposure

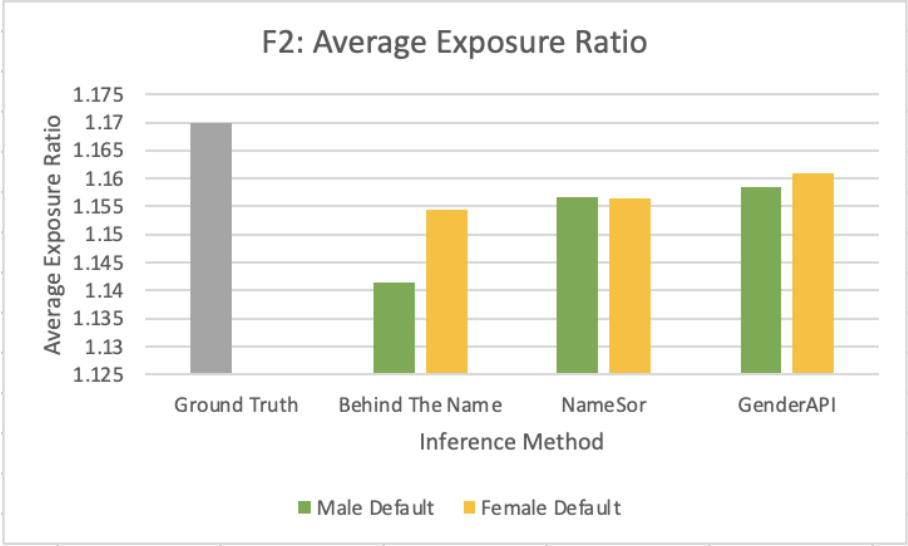


Figure 17: F2, Avg. Exposure Ratio, WNBA/NBA

Figure 17 above show that under a fairness-aware LTR model, the rankings have the lowest average exposure ratio when using the Behind the Name Inference algorithm. As mentioned above the ideal average exposure ratio is equal to 1 meaning that all groups in the ranking have equal exposure or are equally represented.

Across all three inference algorithms, Behind the Name paired with a fairness-aware LTR model produced the lowest average exposure ratio or the most equally represented ranking. However, as shown in the table above, Behind the Name has a 73.4% overall accuracy which is the lowest of the three inference algorithms.

It is also interesting to note that in all three inference algorithms, the average exposure ratio was better or closer to 1 when defaulting to male than when defaulting to female. This means that when the unknowns were defaulted to "male" the groups in the ranking had more equal exposure.

Skew

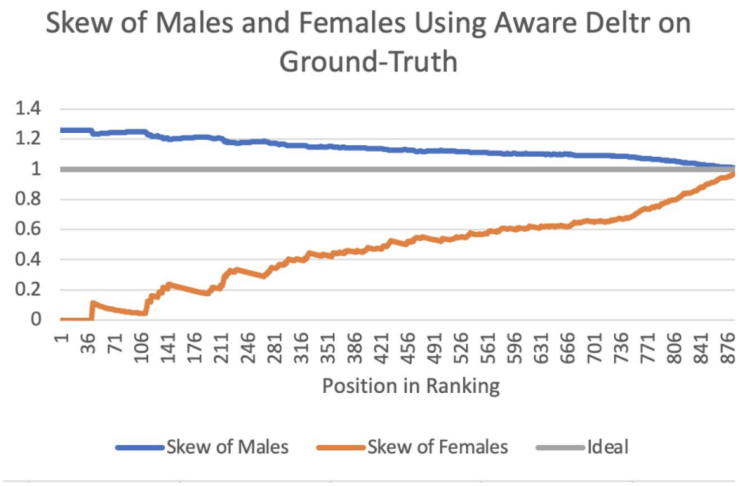
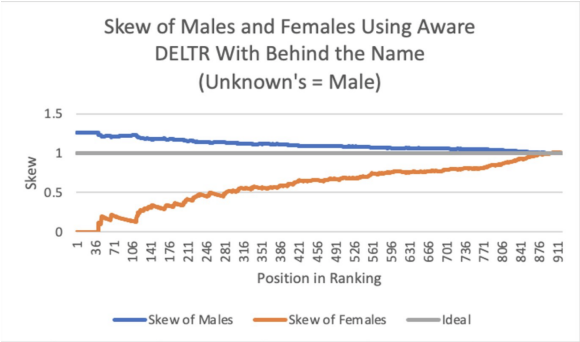
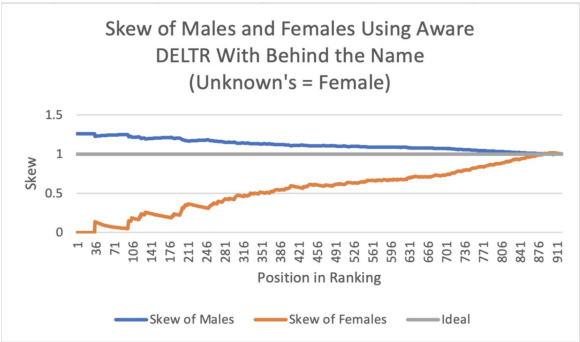


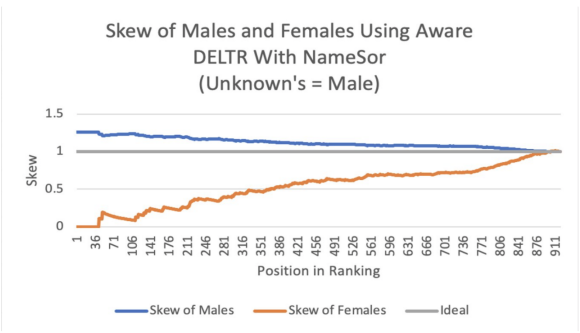
Figure 18: F2, Skew, AD-GT, WNBA/NBA



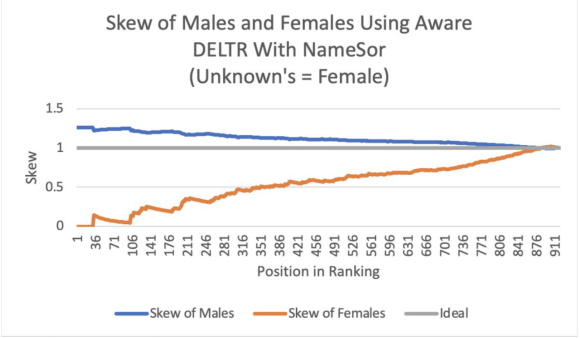
(a) Skew, AD-BTN, Default Male



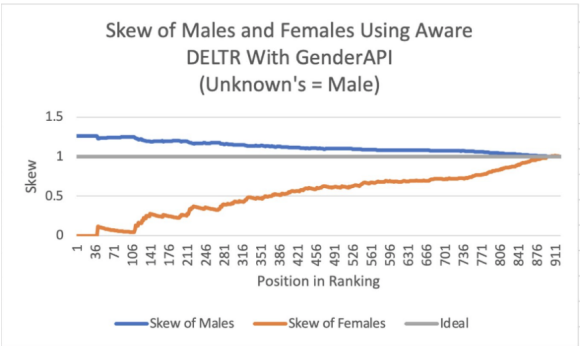
(b) Skew, AD-BTN, Default Female



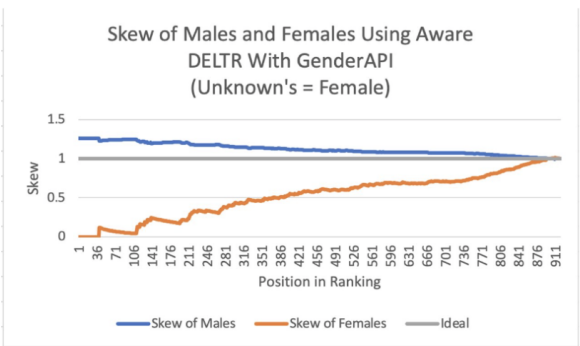
(c) Skew, AD-NSOR, Default Male



(d) Skew, AD-NSOR, Default Female



(e) Skew, AD-GAPI, Default Male



(f) Skew, AD-GAPI, Default Female

Figure 19: F2, Skew AD-INF, WNBA/NBA

Average Positional Difference For Skew

The table below shows the average positional difference of a particular group by comparing two different rankings. The first ranking is produced by the fairness-aware LTR model with ground-truth information and the second is produced by a fairness-aware LTR with inferred demographic information. A positive number means that group experienced an increase in the overall skew across all positions meaning that they became more represented. A negative number means that group experienced a decrease in the overall skew across all positions meaning that they became more under represented.

	Males	Females
AD-BTN (Male Default) ⇒ AD-GT	-0.028344317892543693	0.10905103357604966
AD-BTN (Female Default) ⇒ AD-GT	-0.016309863939465246	0.06275005547236367
AD-NameSor (Male Default) ⇒ AD-GT	-0.013281171490744577	0.051097559788075196
AD-NameSor (Female Default) ⇒ AD-GT	-0.014067279813662199	0.05412200812519515
AD-GAPI (Male Default) ⇒ AD-GT	-0.0121911658332415	0.046903906442629144
AD-GAPI (Female Default) ⇒ AD-GT	-0.009907934448289949	0.03811947411421028

Table 7: F2, Avg. Positional Difference in Skew, WNBA/NBA

By looking at Table 7 and Figure 18, which show skew with ground truth information, and Figure 19, which shows skew with inferred information, it can be seen that the skews of females, the protected group, are generally higher than the skews of males under uncertain or inferred information. This means that under inferred information, a fairness-aware LTR model will increase the representation of the protected group, which in this case is females.

Across the three inference algorithms, there is a notable trend where using the less accurate inference algorithms results in a higher increase in skew for the protected group. Behind the Name, which has a 73.4% inference accuracy rate, had the biggest jump of 0.109. On the other hand, GenderAPI, which has a 92.5% inference accuracy rate, only had a jump of 0.038 in the skew for the protected group.

Conclusion

In the WNBA/NBA experiment, our chosen fairness metrics show that combining a fairness-aware DELTR with inferred demographic information produces more fair rankings than using ground-truth demographic information. More specifically, the rankings were most fair when combining fairness-aware DELTR with Behind the Name, the inference algorithm with the lowest inference accuracy. In addition to this, the rankings were more fair when defaulting the unknowns to male.

Utility Question 2

How does the utility of rankings produced from a fairness-aware LTR model compare when ranking with ground-truth demographic information and when ranking with inferred demographic information?

Average Positional Difference For NDCG

Table 8 shows the average positional difference in the NDCG of two different rankings. The first ranking is produced by a fairness-aware LTR model using ground truth information and the second is produced by a fairness-aware LTR model using inferred demographic information. A positive number means that the NDCG experienced an average increase across all positions meaning that the utility was increased when using a fairness-aware LTR model with inferred demographic information. A negative number means that the NDCG experienced an average decrease in all positions meaning that the ranking lost utility when using a fairness-aware LTR

model with inferred demographic information.

Ranking Comparison	Average Positional Difference
AD-BTN (Male Default) ⇒ AD-GT	0.00005312230911407319
AD-BTN (Female Default) ⇒ AD-GT	-0.00020515713337696797
AD-NameSor (Male Default) ⇒ AD-GT	0.00001293782497761113
AD-NameSor (Female Default) ⇒ AD-GT	-0.00018277905438986072
AD-GAPI (Male Default) ⇒ AD-GT	-0.00001359809513181781
AD-GAPI (Female Default) ⇒ AD-GT	-0.00003584979048720276

Table 8: U2, Avg. Positional Difference in NDCG, WNBA/NBA

Table 8 above shows that when using a fairness aware LTR model on inferred demographic information over ground-truth information, there tends to typically be a loss in utility. Furthermore, there is a greater loss when defaulting the unknown in the inferred demographic information to female or the protected group in this case.

Fairness Question 3
How does the fairness of rankings produced by post-processing a ranking from a fairness-unaware LTR differ when using inferred demographics versus ground-truth demographics?

NDKL

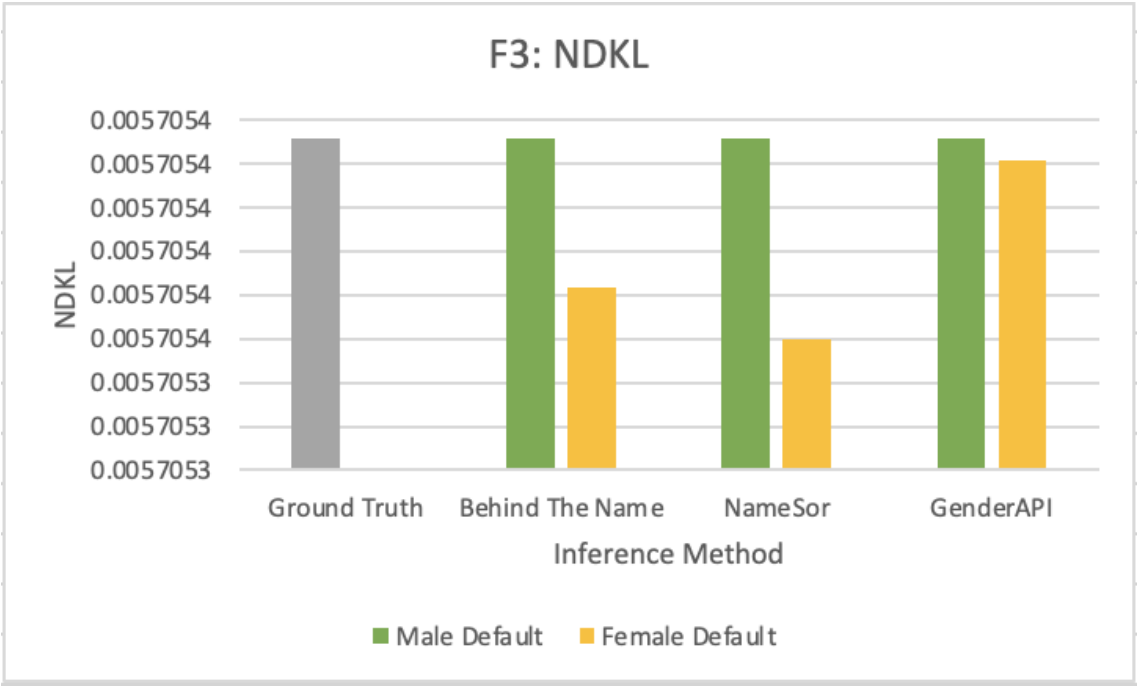


Figure 20: F3, NDKL, WNBA/NBA

By observing Figure 20, where the ideal value of NDKL is 0, meaning that all subgroups in the ranking are proportionally represented, there is a very small difference in the NDKL's when using ground-truth versus inferred demographic information combined with DetConstSort. This is the same across the different inference algorithms used as well as defaulting the unknowns to male or female. In conclusion, NDKL seems to be unaffected in the rankings produced by a post-processing fair ranking algorithm when given ground-truth or inferred demographic information.

Average Exposure Ratio

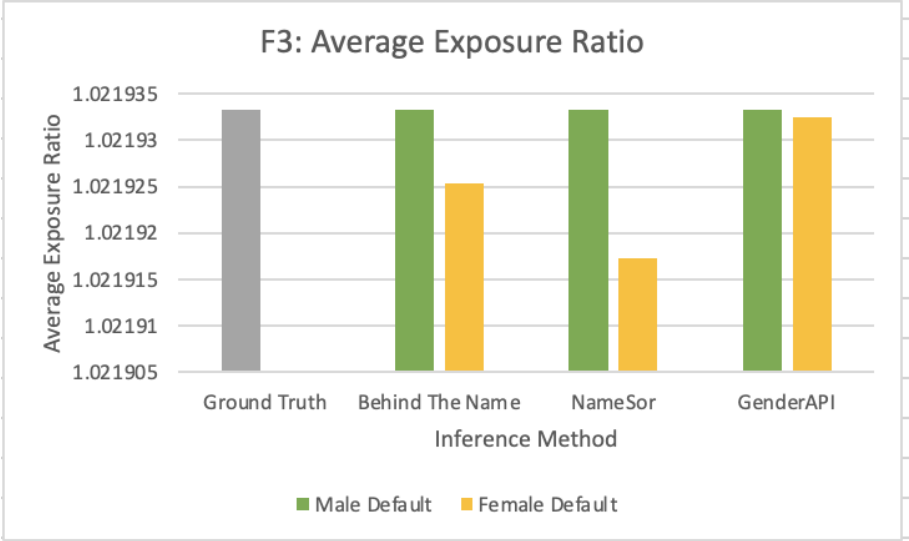


Figure 21: F3, Avg. Exposure Ratio, WNBA/NBA

Figure 21 above shows average exposure ratios across different rankings produced by a post-processing fair ranking algorithm combined with ground truth information versus inferred information. The ideal value is 1, meaning that the groups in the ranking have equal exposure. The graph shows all rankings with a average exposure ratio of close to 1. This means that a post-processing fair ranking algorithm produces rankings that have almost equal exposure when it's combined with ground truth information or inferred information. This stays true across all inference algorithms used in this experiment and when defaulting the unknowns to male or female.

Skew

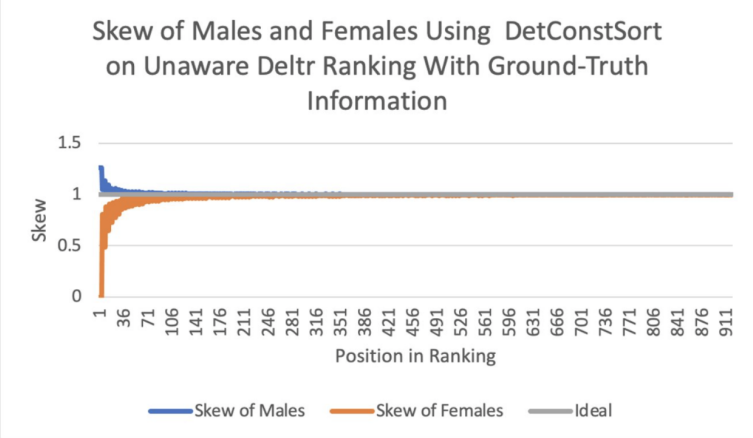
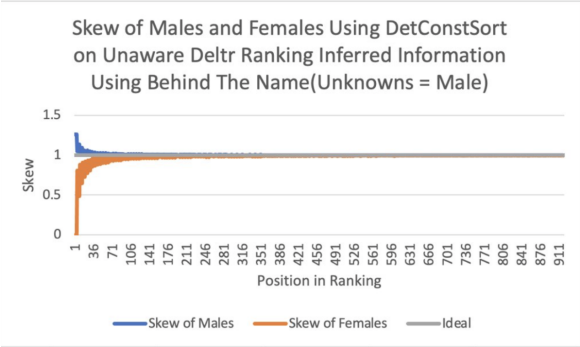
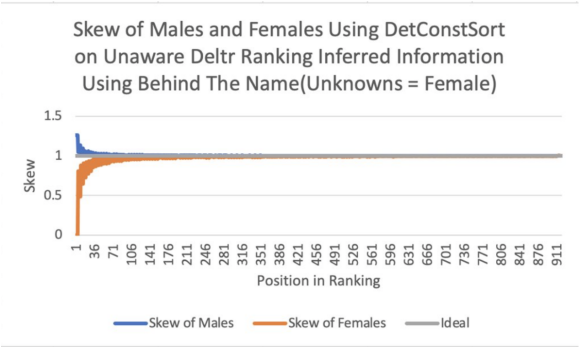


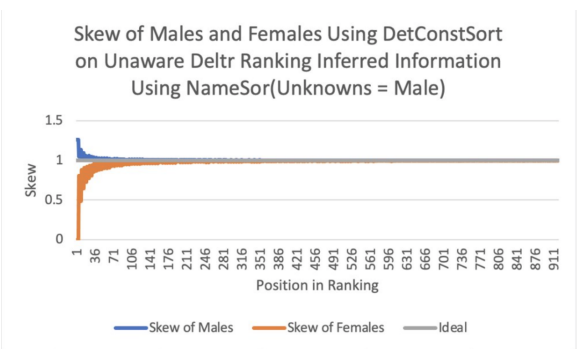
Figure 22: F3, Skew-GT, WNBA/NBA



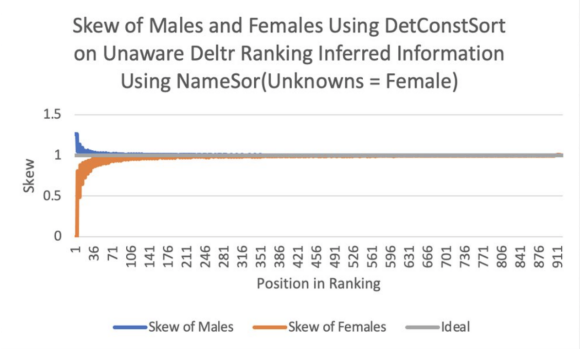
(a) Skew, UD-BTN-DCS, Default Male



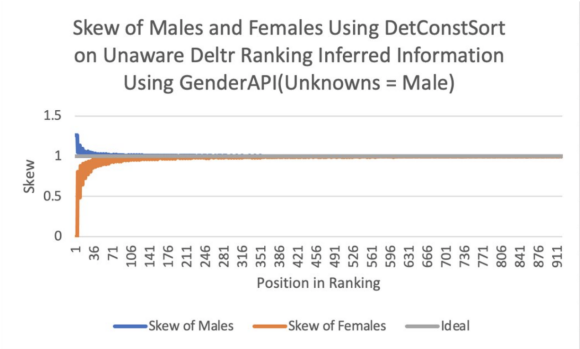
(b) Skew, UD-BTN-DCS, Default Female



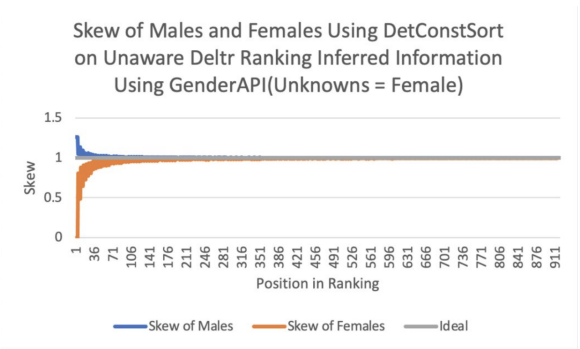
(c) Skew, UD-NSOR-DCS, Default Male



(d) Skew, UD-NSOR-DCS, Default Female



(e) Skew, UD-GAPI-DCS, Default Male



(f) Skew, UD-GAPI-DCS, Default Female

Figure 23: F3, Skew-Inferred, WNBA/NBA

From the skew graphs with inferred information in Figure 23, and with ground truth information in Figure 22, where the ideal value is 1 meaning that all groups are represented equally, there is no difference in skews when using inferred information or ground truth-information combined with a post-processing fair ranking algorithm. This holds true over all inference algorithms and defaulting the unknowns to male or female. For a closer look, consider the average positional difference for skew in Table 9.

Average Positional Difference For Skew

Table 9 shows the average positional difference of a particular group by comparing two different rankings. The first ranking is produced by the post-processing fair ranking algorithm with ground-truth information and the second is produced by the post-processing fair ranking algorithm with inferred demographic information. A positive number means that group experienced an increase in the overall skew across all positions meaning that they became more represented

when using inferred information. A negative number means that group experienced a decrease in the overall skew across all positions meaning that they became more under-represented when using inferred information.

	Males	Females
UD-GT-DCS ⇒ UD-BTN-DCS (Male Default)	0.0	0.0
UD-GT-DCS ⇒ UD-BTN-DCS (Female Default)	-0.0001342796128072806 3	0.0000516623141905919 6
UD-GT-DCS ⇒ UD-NameSor-DCS (Male Default)	0.0	0.0
UD-GT-DCS ⇒ UD-NameSor-DCS (Female Default)	-0.0000269229569293277 7	0.0001035825342912587 8
UD-GT-DCS ⇒ UD-GAPI-DCS (Male Default)	0.0	0.0
UD-GT-DCS ⇒ UD-GAPI-DCS (Female Default)	-0.0000014869446261820 71	0.0000572082379862696

Table 9: F3, Avg Positional Difference in Skew, WNBA/NBA

From the average positional difference for the skew table, it can be seen that the skews did not change significantly when using inferred information over ground truth information combined with a post-processing fair ranking algorithm. It is interesting to note that there was only a change in the skews when the unknowns were defaulted to female when using the inference algorithms in this experiment. In these cases, the females' skew was increased and thus they became more represented when using inferred information and vice versa for the male group.

Conclusion

In the WNBA/NBA experiment, there was no significant impact to the fairness of rankings produced by a post-processing ranking algorithm when combined with ground-truth demographic information versus inferred demographic information. It is important to note that in certain fairness metrics such as NDKL and Skew, the rankings with inferred demographic information where the unknowns were defaulted to female were more fair.

Utility Question 3

How does the utility of rankings produced by post-processing a ranking from a fairness-unaware LTR differ when using inferred demographics versus ground-truth demographics?

Average Positional Difference For NDCG

Table 10 shows the average positional difference in the NDCG of two different rankings. The first ranking is produced by a post-processing fairness aware ranking algorithm using ground truth information and the second is produced by a post-processing fairness aware ranking algorithm using inferred demographic information. A positive number means that the NDCG

experienced an average increase across all positions meaning that the utility was increased when using inferred demographic information. A negative number means that the NDCG experienced an average decrease in all positions meaning that the ranking lost utility when using inferred demographic information.

Ranking 1 ⇒ Ranking 2	Average Positional Difference in NDCG
UD-GT-DCS ⇒ UD-BTN-DCS (Male Default)	0.0001238366582778633
UD-GT-DCS ⇒ UD-BTN-DCS (Female Default)	-0.0001938184252842118
UD-GT-DCS ⇒ UD-NameSor-DCS (Male Default)	0.00005285976856808245
UD-GT-DCS ⇒ UD-NameSor-DCS (Female Default)	-0.00018155085707855407
UD-GT-DCS ⇒ UD-GAPI-DCS (Male Default)	0.00002122845368936974
UD-GT-DCS ⇒ UD-GAPI-DCS (Female Default)	-0.000014899381531695915

Table 10: U3, Avg. Positional Difference in NDCG, WNBA/NBA

Based on Table 10, there was no significant change in the utility of rankings when using inferred demographic information over ground truth information in combination with the post-processing fair ranking algorithm. However, it is important to note that the utility increased with inferred information across all inference algorithms when the unknowns were defaulted to male. On the other hand, the utility decreased slightly with inferred information across all inference algorithms when the unknowns were defaulted to female.

Fairness Question 4

How does the fairness of rankings obtained using ground-truth demographics differ from a fairness-aware in-processing LTR model to a post-processing fair ranking algorithm?

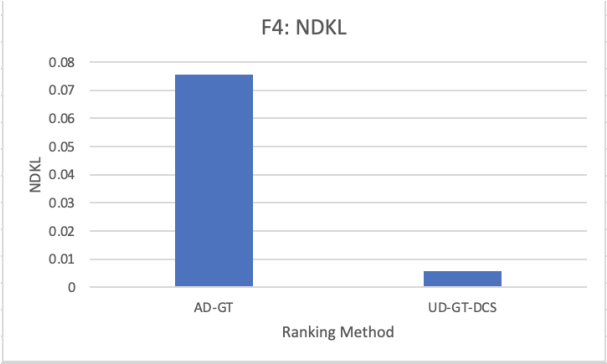


Figure 24: F4, NDKL, WNBA/NBA

Looking at the NDKL graph above in figure 24 where the ideal value of NDKL is 0 meaning that all subgroups in the ranking are proportionally represented, UD-GT-DCS is very close to 0 as opposed to AD-GT which is at 0.07. This means that under ground truth demographic information, a post-processing ranking is potentially more fair than an in-processing fairness-aware LTR model in terms of representing subgroups proportionally.

Average Exposure Ratio

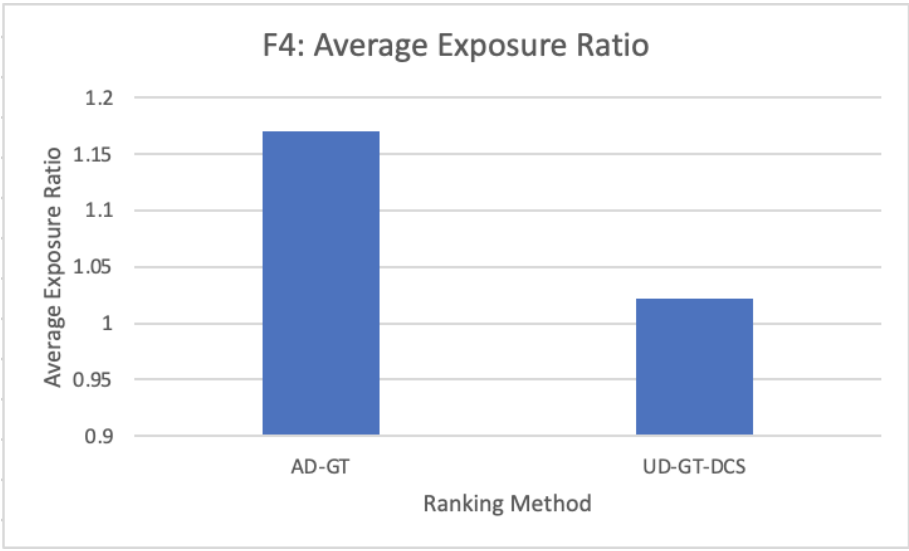
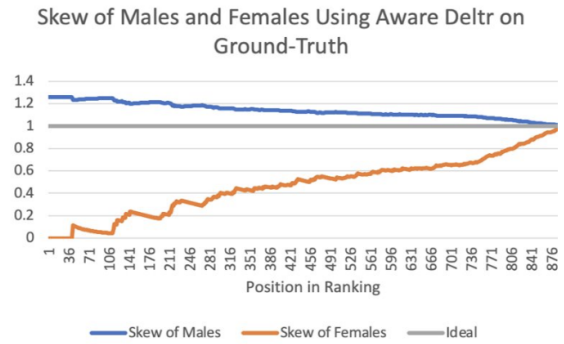


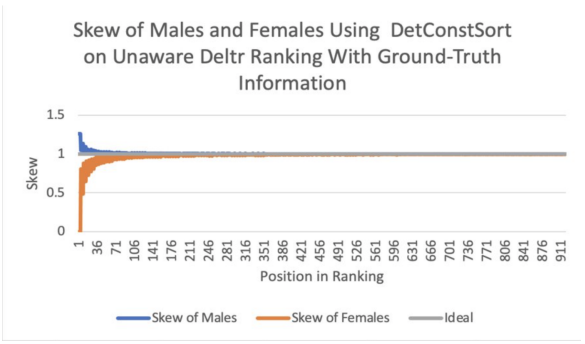
Figure 25: Average Exposure Ratio Graph for Fairness Question 4, WNBA/NBA

Looking at the average exposure ratio bar graph in figure 25above where the ideal value is one meaning that all groups are represented equally, UD-GT-DCS is closer to 1 than AD-GT. This means that under ground truth demographic information a post-processing ranking is potentially more fair than in-processing fairness-aware LTR model in terms of equal exposure over all subgroups of ranking.

Skew



(a) Skew, AD-GT



(b) Skew, UD-GT-DCS

Figure 26: Skew of Rankings Using Ground Truth InformationComparing Post-Processing Ranking Algorithm with Fairness-Aware LTR Model

Looking at the skew graphs above in figure 26 where the ideal value is 1 meaning both subgroups are represented equally, the ranking UD-GT-DCS converges to 1 much faster than AD-GT. This means that under ground truth information a post-processing ranking is potentially more fair than an in-processing fairness-aware LTR model in terms of representing different subgroups proportionally.

Average Positional Difference For Skew

Table 11 shows the average positional difference of a particular group by comparing two different rankings. The first ranking is produced by the fairness-aware LTR model with ground-truth information and the second is produced by a post-processing ranking algorithm fairness-aware LTR also with ground truth information. A positive number means that group experienced an increase in the overall skew across all positions meaning that they became more represented. A negative number means that group experienced a decrease in the overall skew across all positions meaning that they became more under represented.

	Males	Females
AD-GT ⇒ UD-GT-DCS	-0.12741905850355725	0.49022806192684315

Table 11: F4, Avg. Positional Difference in Skew, WNBA/NBA

Table 11 shows that a post-processing ranking algorithm was on average able to increase the overall skew or representation of the female group by 0.49 which was the protected group in this case. In addition to this, the males, which were the over-represented group, showed a lower representation when ranked with the post processing ranking algorithm.

Conclusion

For the WNBA/NBA experiment, our chosen fairness metrics showed that, under ground-truth demographic information, the post-processing fair ranking algorithm produced more fair rankings than the in-processing fairness-aware DELTR.

Utility Question 4 *How does the utility of rankings obtained from a fairness-aware in-processing LTR model, ranked using ground-truth demographics, compare to the utility of rankings obtained from post-processing a ranking from a fairness-unaware LTR?*

Average Positional Difference For NDCG

Table 12 shows the average positional difference in the NDCG of two different rankings. The first ranking is produced by an in-processing fairness-aware LTR model using ground truth information and the second is produced by a post-processing ranking algorithm also using ground truth information. A positive number means that the NDCG experienced an average increase across all positions, meaning that the utility was increased when using a post-processing algorithm with ground truth demographic information. A negative number means that the NDCG experienced an average decrease in all positions meaning that the ranking lost utility when using a post-processing ranking algorithm with ground truth demographic information.

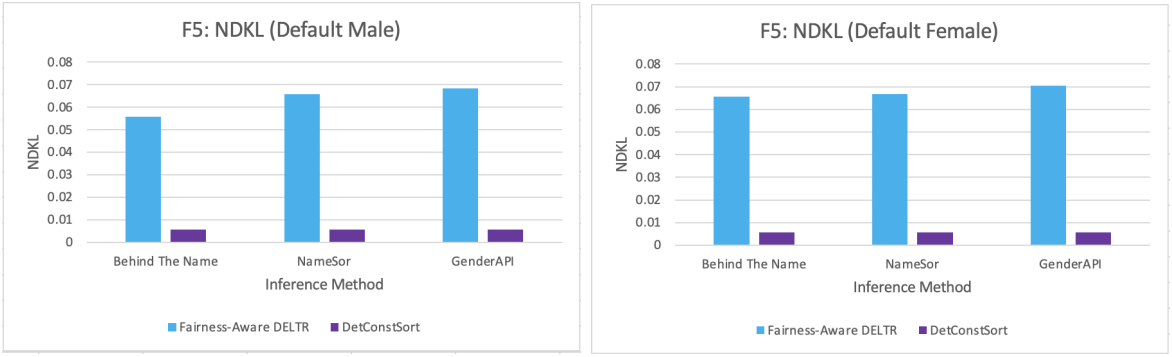
Ranking Comparison	Average Positional Difference
AD-GT ⇒ UD-GT-DCS	0.31924590647138795

Table 12: U4, Avg. Positional Difference in NDCG, WNBA/NBA

Table 12 shows that the utility of a ranking increases by an average of 0.319 across all positions when using ground truth demographic information with a post-processing ranking algorithm as opposed to an in-processing fairness aware LTR model.

Fairness Question 5 *How does the fairness of rankings obtained from a fairness-aware in-processing LTR model, ranked using inferred demographics, compare to the fairness of rankings obtained from a post-processing fair ranking algorithm, ranked using inferred demographics?*

NDKL



(a) F5, NDKL, Male Default, WNBA/NBA (b) F5, NDKL, Female Default, WNBA/NBA

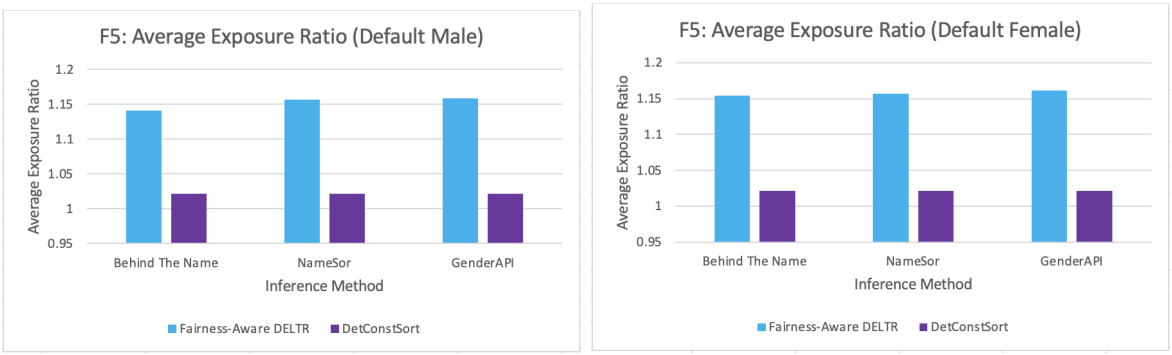
Figure 27: F5, NDKL, WNBA/NBA

When using inferred demographics, according to the graphs above, a post-processing fair ranking algorithm achieves a much lower NDKL than an in-processing fairness-aware LTR model. This means that the post-processing ranking has subgroups that are more proportionally represented and therefore potentially more fair.

The post-processing fairness-aware LTR model has a much lower NDKL no matter which inference algorithm used but when using the in-processing LTR model, the NDKL seems to decrease as the accuracy of the inference algorithm used decreases. For example the in-processing method combined with Behind the Name, which has the lowest inference accuracy, also has the lowest NDKL. On the other hand the in-processing method combined with GenderAPI, which has the highest inference accuracy, has the highest NDKL.

The NDKL across all inference algorithms appears to be higher when defaulting to female and using the in-processing LTR model.

Average Exposure Ratio



(a) F5, Avg. Exp Ratio, Male Default, WNBA/NBA (b) F5, Avg. Exposure Ratio, Female Default, WNBA/NBA

Figure 28: F5, Avg. Exposure Ratio, WNBA/NBA

Looking at the average exposure ratio graphs in figure 28a and 28b where the ideal value is 1 meaning that all groups in the ranking have an equal amount of exposure, the DetConstSort values consistently have a much better average exposure ratio than their AD counterparts. This means that under inferred demographic information, the post-processing fair ranking algorithm overall produces rankings where the subgroups have more equal exposure than the in-processing fairness-aware LTR model.

For the post-processing method, all the inference algorithms performed relatively the same and were around 1.02. However, for the in-processing fairness-aware LTR model, Behind the Name had an average exposure ratio that was closer to 1 than when using the other inference algorithms.

Similar to above, the post-processing method did not alter when defaulting the unknowns to male or female when using inference algorithms. However, the in-processing model combined with the Behind the Name Inference algorithm showed a more ideal value of average exposure when defaulting to male.

Skew

Behind the Name

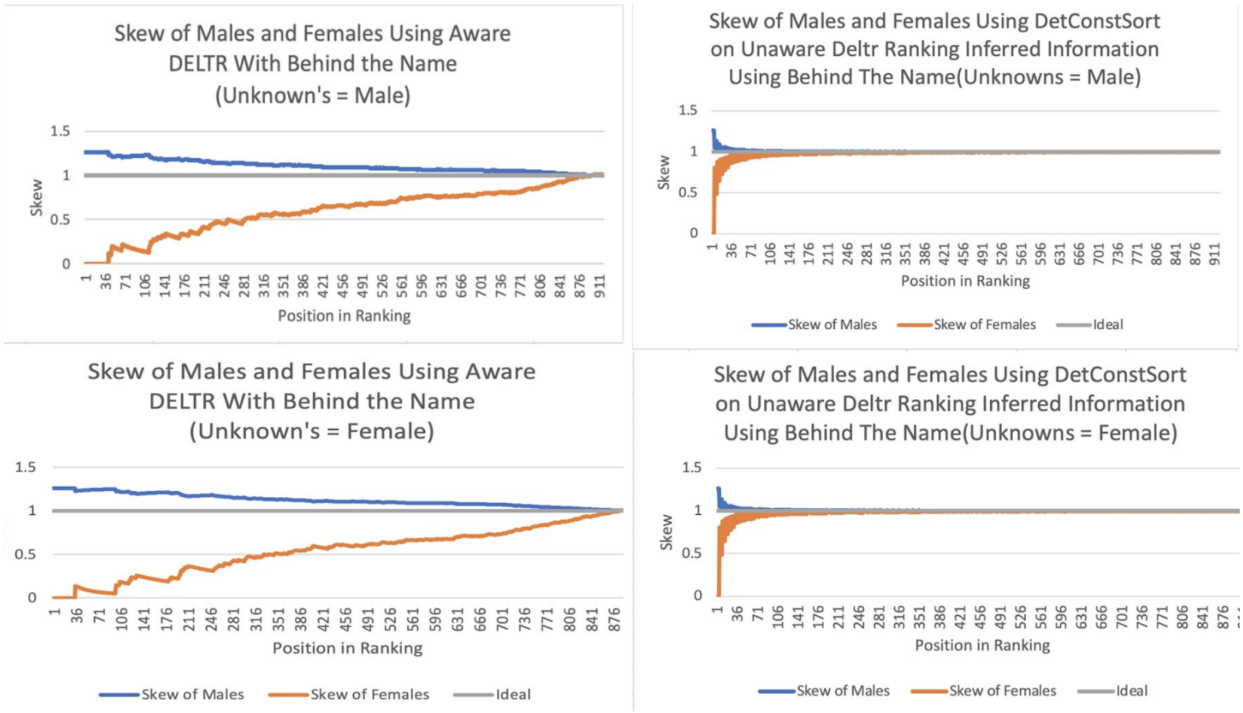


Figure 29: F5, Skew-BTN, WNBA/NBA

NameSor

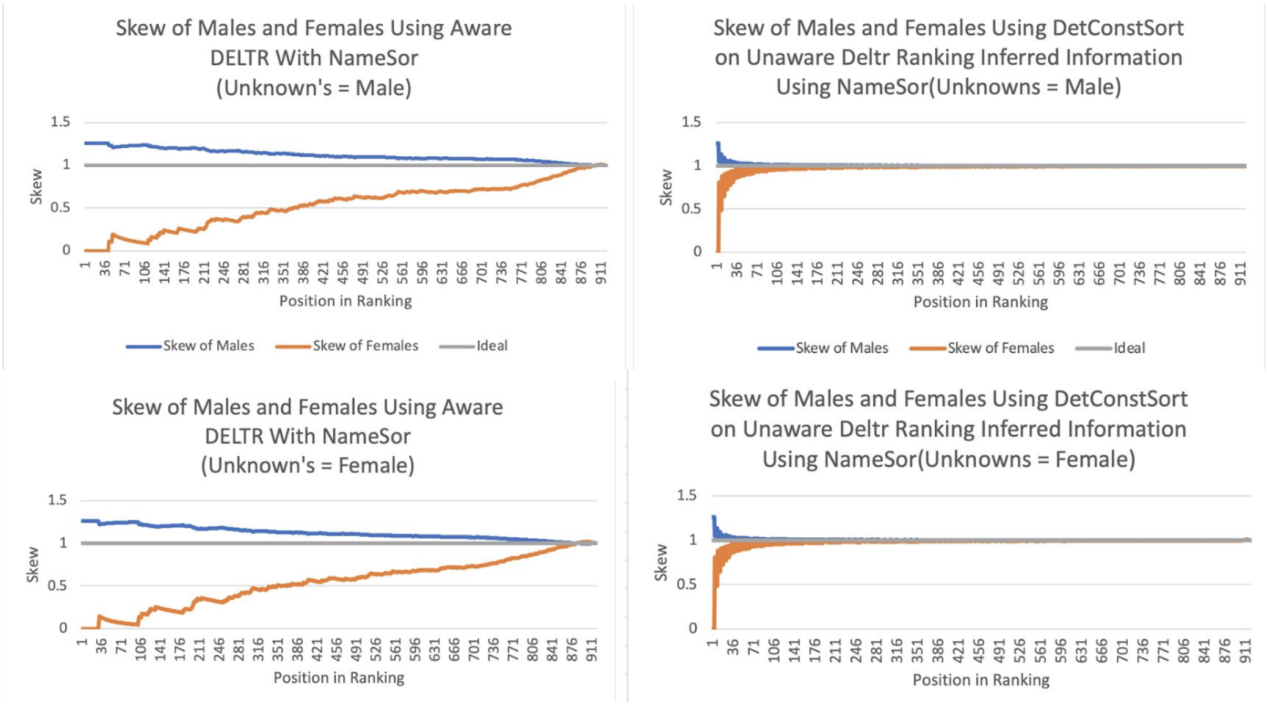


Figure 30: F5, Skew-NSOR, WNBA/NBA

GenderAPI

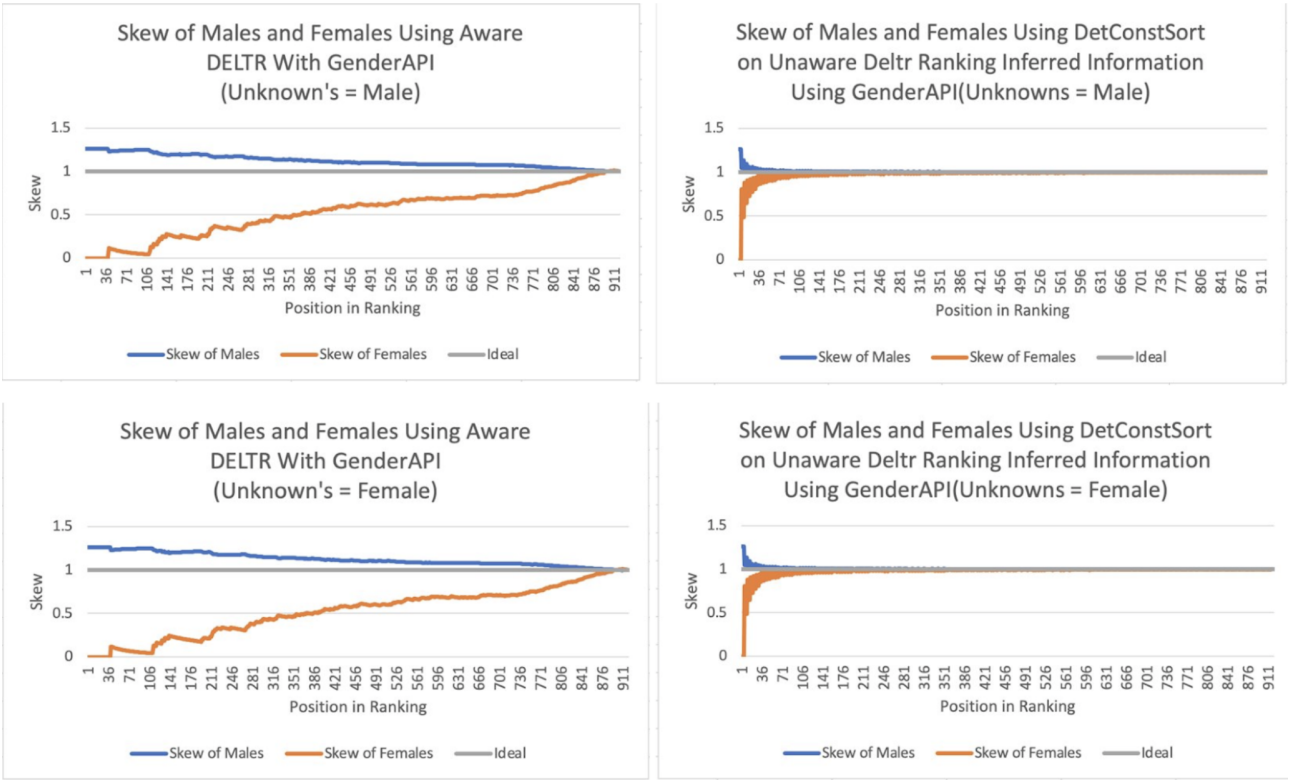


Figure 31: F5, Skew-GAPI, WNBA/NBA

From the skew graphs above where the ideal value is 1, it can clearly be seen that no matter the inference algorithm used the skews of males and females converge faster to 1 under the post-processing fair ranking algorithm DetConstSort. On the other hand the fairness-aware

LTR model, DELTR, produces skews that show females being underrepresented for a majority of the ranking (a value under 1) and the males being over represented (a value of over 1).

Average Positional Difference For Skew

Table 13 shows the average positional difference of a particular group by comparing two different rankings. The first ranking is produced by the fairness-aware LTR model with inferred information and the second is produced by a post-processing fair ranking algorithm also with inferred information. A positive number means that group experienced an increase in the overall skew across all positions meaning that group became more represented. A negative number means that the group experienced a decrease in the overall skew across all positions meaning that group became less represented.

Ranking 1 ⇒ Ranking 2	Males	Females
AD-BTN (Default Male) ⇒ UD-BTN-DCS (Default Male)	-0.09907474061101322	0.3811770283507928
AD-BTN (Default Female) ⇒ UD-BTN-DCS (Default Female)	-0.11112262252537247	0.4275296687686694
AD-NameSor (Default Male) ⇒ UD-NameSor-DCS (Default Male)	-0.11413788701281255	0.4391305021387683
AD-NameSor (Default Female) ⇒ UD-NameSor-DCS (Default Female)	-0.11337870164682394	0.4362096363359385
AD-GAPI (Default Male) ⇒ UD-GAPI-DCS (Default Male)	-0.11522789267031552	0.4433241554842131
AD-GAPI (Default Female) ⇒ UD-GAPI-DCS (Default Female)	-0.1175126109998933	0.45211430863643093

Table 13: F5, Avg. Positional Difference in Skew, WNBA/NBA

Looking at the average positional difference for skew table above it seems as the accuracy of the inference algorithm increased, the representation of the protected group, females, increased when comparing DELTR, a fairness-aware LTR model, to DetConstSort, a post-processing fair ranking algorithm.

Across all three algorithms, when defaulting the unknowns to male or female, the post-processing ranking method proved to increase representation of the protected group, females, and lower the representation of the over-represented group, males.

Conclusion

For the WNBA/NBA experiment, our chosen metrics show that rankings produced by DetConstSort, the post-processing fair ranking algorithm, produces more fair rankings than

the in-processing fairness-aware DELTR. This remains true across all inference methods and whether we defaulted the unknowns to male or female.

Utility Question 5 *How does the utility of rankings obtained from a fairness-aware in-processing LTR model, ranked using inferred demographics, compare to the utility of rankings obtained from a post-processing fair ranking algorithm, ranked using inferred demographics?*

Average Positional Difference For NDCG Table 14 the average positional difference in the NDCG of two different rankings. The first ranking is produced by an in-processing fairness-aware LTR model using inferred information and the second is produced by a post-processing fair ranking algorithm also using inferred information. A positive number means that the NDCG experienced an average increase across all positions, meaning that the utility was increased when using a post-processing algorithm with inferred demographic information. A negative number means that the NDCG experienced an average decrease in all positions meaning that the ranking lost utility when using a post-processing ranking algorithm with inferred information.

Ranking 1 ⇒ Ranking 2	Average Positional Difference in NDCG
AD-BTN (Default Male) ⇒ UD-BTN-DCS (Default Male)	0.31931662082055123
AD-BTN (Default Female) ⇒ UD-BTN-DCS (Default Female)	0.31925724517948006
AD-NameSor (Default Male) ⇒ UD-NameSor-DCS (Default Male)	0.3192858284149784
AD-NameSor (Default Female) ⇒ UD-NameSor-DCS (Default Female)	0.31924713466869803
AD-GAPI (Default Male) ⇒ UD-GAPI-DCS (Default Male)	0.3192807337345897
AD-GAPI (Default Female) ⇒ UD-GAPI-DCS (Default Female)	0.319266856880343

Table 14: U5, Avg. Positional Difference in NDCG, WNBA/NBA

The average positional difference for NDCG chart above shows that under inferred information the utility of rankings increases across all inference algorithms when using a post-processing fair ranking algorithm over an in-processing fairness-aware LTR model.

4.2 2017 Boston Marathon Experiment Results

Based on the methodology described above, we ran the experiment on a dataset consisting of runners who raced in 2017 Boston Marathon. For the purposes of this experiment only times at 3 different stages (5K,15k, Half), Pace, Name, M/F and Official Time were used.

Data Column	Description
Split Times (5k, 15k, Half)	Used as additional training columns.
Pace	Used as an additional training column.
Name	The name column was used to infer the gender attribute.
M/F	This column specifies the gender of each participant and is used as the protected attribute.
Official Time	The over all, official time for each participant and is used as the score attribute.

Table 15: Boston Marathon Dataset Columns.

4.2.1 Cleaning and Splitting Boston Marathon

The Boston Marathon dataset was cleaned and ranked according to the score attribute, Official Time, in ascending order. The dataset was then split with 80/20 train/test split, as was decided in the methodology.

4.2.2 Training With Boston Marathon Data

Using the 80% train split of the Boston Marathon Dataset, we trained a fairness-aware DELTR model using gamma = 1.0 and a fairness-unaware DELTR model using gamma = 0.0.

4.2.3 Inferring Boston Marathon Data

According to the methodology described above, Using the 20% test split we developed inferred datasets with inferred genders for the Boston Marathon race participants. We used the following inference algorithms:

1. Behind The Name
2. NameSor
3. GenderAPI

When inferring the test data on each inference algorithm there was a male default inferred result and a female default inferred result. This is because, at times the inference algorithms were not able to produce a predicted gender from a name, in these cases we defaulted all the unknowns to male for one data set and all the unknowns to female in another data set. Gender API is the least accurate at inferring gender while NameSor is the most accurate. Although, Behind the Name inference has a 98.15% accuracy, it does not identify a very high amount of datapoint. This can be seen the Table 32 and Figure 33 below.

	Percent Unidentifiable	Accuracy Excluding Unidentifiable Results	Accuracy Including Unidentifiable Results	% Female in Unidentified	% Male in Unidentified	% Accuracy when default is Male	% Accuracy when default is Female
Behind the Name	12.96%	98.15%	79.54%	45.79%	54.21%	50.20%	89.82%
NameSor	11.08%	99.10%	88.12%	59.46%	40.54%	94.71	92.61
GenderAPI	3.89%	96.26%	92.51%	38.46%	61.54%	94.91%	94.91%

Figure 32: Boston Marathon Inferred Gender Accuracy.

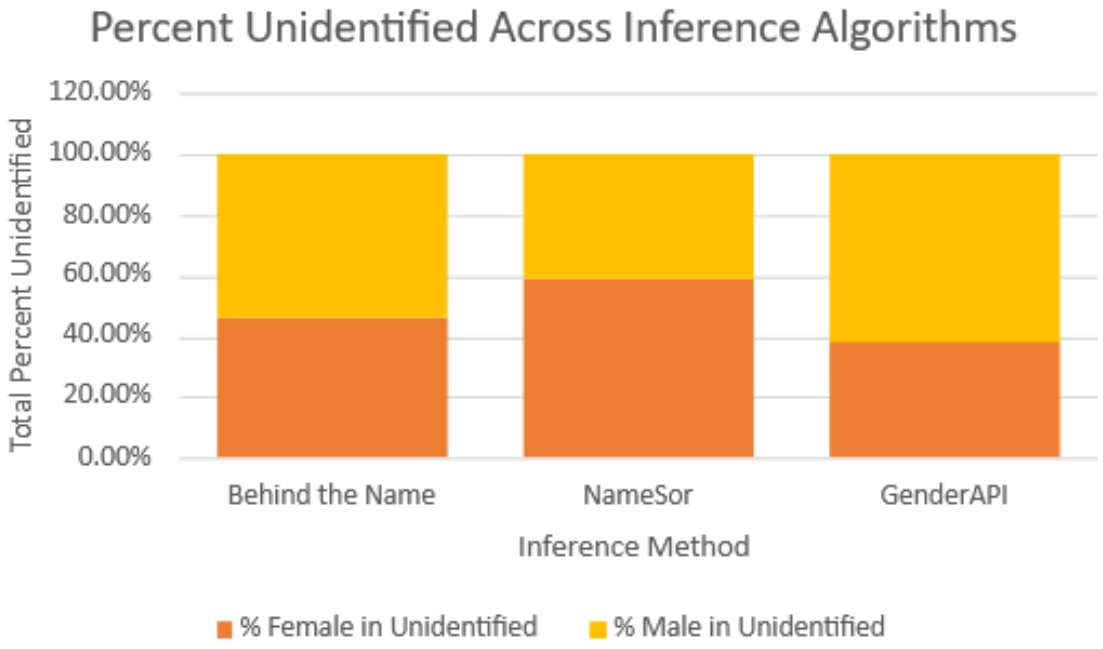


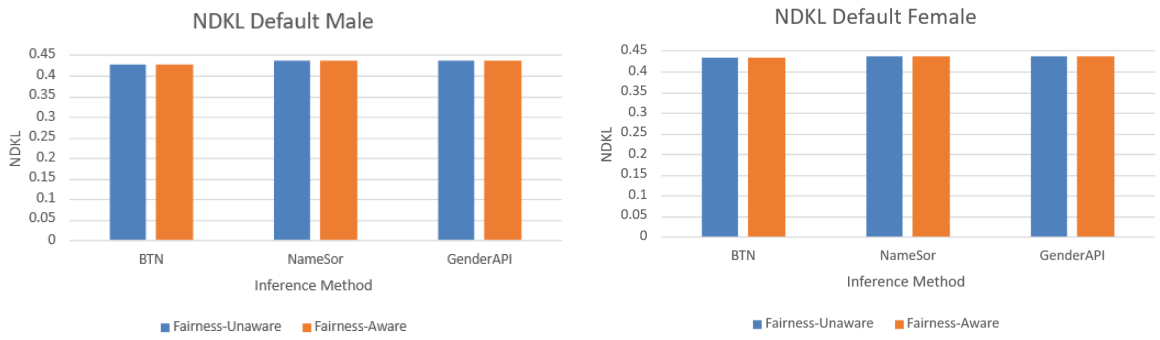
Figure 33: Percent Unidentified Across Inference Algorithms for Default Male and Default Female.

4.2.4 Research Questions

In this section we explore the different fairness and utility research questions described in the methodology with respect to the Boston Marathon data set.

Fairness Question 1
Given uncertain demographic information, how does the fairness of rankings produced from a fairness-unaware LTR model compare to the fairness of rankings produced from a fairness-aware LTR model?

NDKL



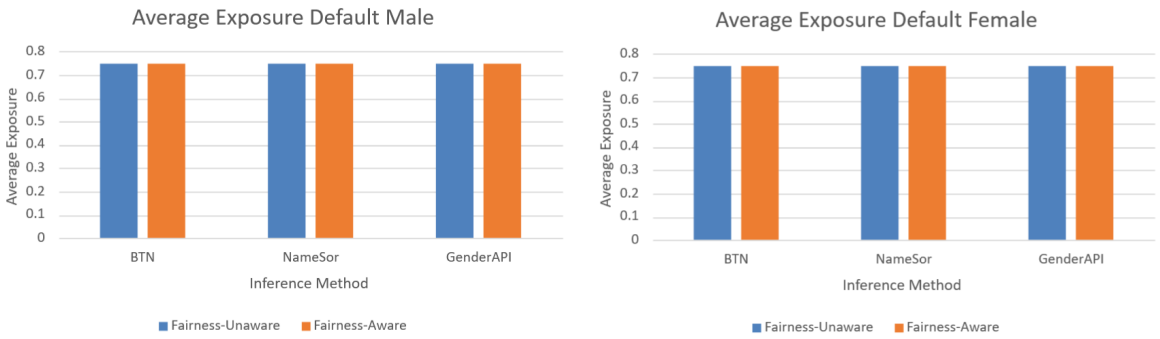
(a) NDKL default male inference.

(b) NDKL default female inference.

Figure 34: NDKL for Fairness-Aware DELTR Using Inference Algorithms

The Figures 34a and 34b above visualize the change in NDKL over all three inference algorithms comparing Fairness-Unaware and Fairness-Aware DELTR. There is some change across inference algorithms in Default Male, most notably with Behind the Name having significantly more fair results. Although there differences between inference algorithms, there is little difernce between fairness-unaware and fairness aware DELTR. According to the NDKL results, there is not fairness difference between the two DELTR models when the dataset is ranked, regardless of the inference method.

Average Exposure Ratio



(a) Average Exposure Ratio with Default Male inference.

(b) Average Exposure Ratio with Default Female inference.

Figure 35: Average Exposure Ratio for Fairness-Aware DELTR Using Inference Algorithms

Figures 35a and 35b show the results found when calculating Average Exposure Ratio of males over females. If the Average Exposure Ratio is less than one then males are under-represented and if it is greater than one then males are over represented. Therefore over the three inference methods males are consistently under-represented regardless of the gamma value used to rank them. When looking at Behind the Name method, the Average Exposure Ratio value is closer to one, more so than NameSor or GenderAPI.

Average Positional Difference in Skew

Figure 36 shows the average positional difference of a particular group by comparing two different rankings. The first ranking is produced by the fairness-unaware DELTR and the second is produced by the fairness-aware DELTR. A positive number means that group experienced an increase in the overall skew across all positions meaning that they became more represented. A negative number means that group experienced a decrease in the overall skew across all positions meaning that they became more under represented. When looking at skew, we decided that the Average Positional Difference in Skew provided clear information, as opposed to line graphs of the skew.

Ranking 1 => Ranking 2	**Males**	**Females**
UD-BTN (Male Default) => AD-BTN (Male Default)	-2.901174395395261e-06	2.901174395395261e-06
UD-BTN (Female Default) => AD-BTN (Female Default)	-3.526515872847969e-06	3.526515872847969e-06
UD-NameSor (Male Default) => AD-NameSor (Male Default)	-3.6826715572545003e-06	3.6826715572546663e-06
UD-NameSor (Female Default) => AD-NameSor (Female Default)	-3.5770752402006917e-06	3.577075240200304e-06
UD-BTN (Male Default) => AD- GAPI (Male Default)	-3.675896607247031e-06	3.6758966072411587e-06
UD-BTN (Female Default) => AD- GAPI (Female Default)	-3.622516247771717e-06	3.6225162477627445e-06

Figure 36: Average Positional Difference in Skew using Inference Algorithms.

The results show that there is a very slight increase in the skew for the female group. Although, because the average positional skew difference is so small, there are no significant changes in the representation of the female group from the fairness-aware and fairness-unaware models.

Results collected from the Boston Marathon dataset show little to no differences in fairness between rankings using fairness-aware and fairness-unaware models. There is a slight improvement in fairness when using Behind the Name inference over using NameSor and GenderAPI.

Conclusion

The overall findings of Fairness Question 1 found that Behind the Name performed consistently more fair than NameSor or GenderAPI. This improved fairness performance in Behind

the Name comes at the cost of overall accuracy in predicting the gender during inference. The performance of fairness between fairness-unaware and fairness-aware DELTR models is negligible.

Utility Question 1
Given uncertain demographic information, how does the utility of rankings produced from a fairness-unaware LTR model compare to the utility of rankings produced from a fairness-aware LTR model?

Table 37 shows the average positional difference in NDCG values for the different rankings produced from fairness-unaware and fairness-aware DELTR models. A positive NDCG value indicates a better performance in utility of the ranking. While a negative NDCG value will prove a ranking has a decrease in utility.

Ranking Comparison	Average Positional Difference
UD-BTN (Male Default) => AD-BTN (Male Default)	-0.00060641666690225062
UD-BTN (Female Default) => AD-BTN (Female Default)	-0.00061270644768928 42
UD-NameSor (Male Default) => AD-NameSor (Male Default)	-0.0006121810196060364
UD-NameSor (Female Default) => AD-NameSor (Female Default)	-0.0006127662243328079
UD-GAPI (Male Default) => AD- GENDERAPI (Male Default)	-0.0006121810196060364
UD- GAPI (Female Default) => AD- GAPI (Female Default)	-0.0006127662243328079

Figure 37: Average Positional Difference in NDCG using Inference Algorithms.

The results from Table 37 will show that across all inference algorithms, the rankings all lost utility value when comparing UD-INF to AD-INF. This is expected because the fairness-unaware DELTR models will learn solely on the utility value, while aware-DELTR models will try and balance fairness with utility.

The overall finding of Utility Question 1 shows the expected decrease in utility value when a fairness-aware DELTR model is used.

Fairness Question 2
How does the fairness of rankings produced from a fairness-aware LTR model compare when ranking with ground-truth demographic information and when ranking with inferred demographic information?

NDKL

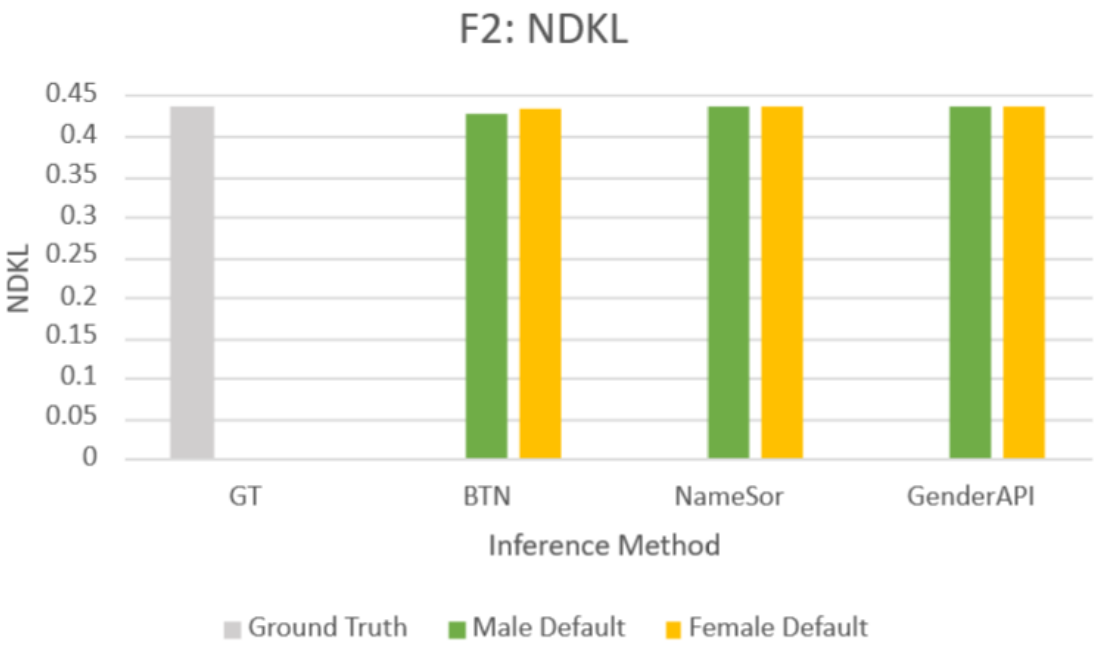


Figure 38: NDKL using Ground Truth and Aware-DELTR with Inference.

The results from Figure 38 shows that all three inference methods performed with a NDKL value within 0.01 of the Ground Truth value. The inference method did perform slightly more fairly, having a value closer to zero, than the Ground Truth ranking.

Average Exposure Ratio

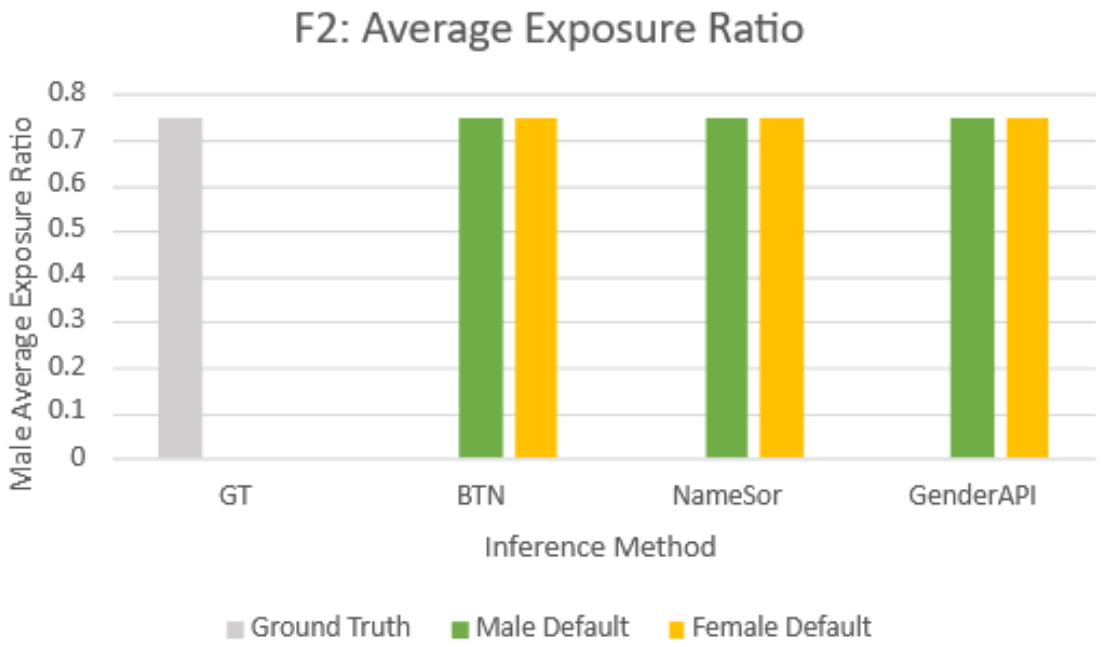


Figure 39: Average Exposure Ratio using Ground Truth and Aware-DELTR with Inference.

Figure 39 shows that the Average Exposure Ratios of the inference methods had values slightly closer to one than the ground truth. The inference methods did not seem to have significance in the results with compared to the ground truth. The results also show that males are under-represented in this ranking and females are over-represented.

Average Positional Difference in Skew

Table 16 shows the average positional difference in skew when comparing fairness-aware DELTR in combination with ground-truth demographics versus inferred demographics. A positive value indicates that the specified group's average representation in the ranking increased when using inferred demographics over ground-truth demographics. The value is negitive for the female group, which means that they lost exposure in the top rankings. This ranking was less fair.

Ranking 1 => Ranking 2	Males	Females
AD-GT => AD-BTN (Male Default)	0.011922862995600294	-0.01192286299560028
AD-GT => AD-BTN (Female Default)	0.0025575138115213644	-0.0025575138115213666
AD-GT => AD-NameSor (Male Default)	0.002059349862925862	-0.002059349862925859
AD-GT => AD-NameSor (Female Default)	0.0018200641467584105	-0.0018200641467584094
AD-GT => AD- GAPI (Male Default)	0.0015912217834144653	-0.0015912217834144634
AD-GT => AD- GAPI (Female Default)	0.0008408276927160552	-0.0008408276927160559

Table 16: Average Positional Difference in Skew using Ground Truth and Aware-DELTR with Inference.

Looking at the results in Table 16 the average positional difference in skew there is a decrease in visibility of females from the ground truth to the inferred rankings.

Conclusion

The results of Fairness Question 2 show that the inference methods produce more fair results than the ground truth ranking. There is only a marginal difference between using default male or default female in the inference methods.

Utility Question 2

How does the utility of rankings produced from a fairness-aware LTR model compare when ranking with ground-truth demographic information and when ranking with inferred demographic information?

Average Positional Difference in NDCG

Table 16 shows the average positional difference in NDCG when comparing fairness-aware DELTR in combination with ground-truth demographics versus inferred demographics. A positive value indicates that there was an increase in the average utility of the ranking when using inferred demographic information over ground-truth and a negative value means a decrease in average utility.

Ranking Comparison	Average Positional Difference
AD-GT => AD-BTN (Male Default)	-0.03927754174696352
AD-GT => AD-BTN (Female Default)	0.0002924793929961705
AD-GT => AD-NameSor (Male Default)	-0.002981775154304716
AD-GT => AD-NameSor (Female Default)	0.000686354617430894
AD-GT => AD- GAPI (Male Default)	-0.002685142501530737
AD-GT => AD- GAPI (Female Default)	-0.000876277407625316

Table 17: Average Positional Difference in NDCG using Ground Truth and Aware-DELTR with Inference.

In Table 17 above, the overall utility when using DELTR rankings improved across Behind the Name and NameSor algorithms when Female is default. The utility value decreases in both male and female default GenderAPI, this can be seen by the negative Average Positional Difference in NDCG.

Fairness Question 3
How does the fairness of rankings produced by post-processing a ranking from a fairness-unaware LTR differ when using inferred demographics versus ground-truth demographics?

NDKL

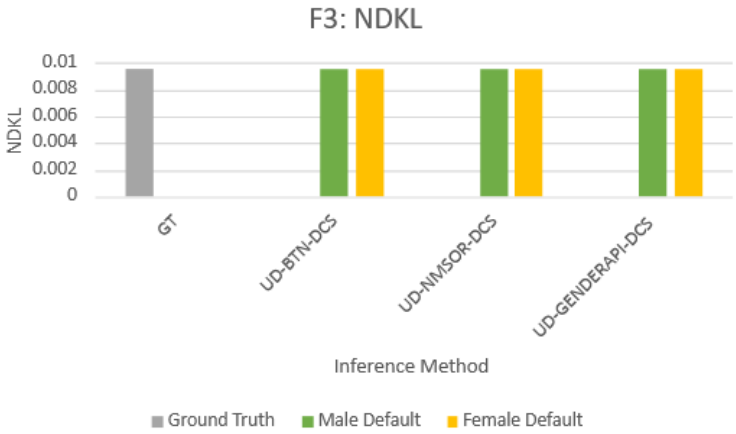


Figure 40: NDKL using UD-GT-DCS and UD-INF-DCS.

The findings from Figure 40 compares the NDKL values of rankings produced by Unaware DELTR trained on Ground Truth values and ranked with DetConstSort, to the rankings of Unaware DELTR trained on the three inference algorithms and DetConstSort. The NDKL values have no change across the different inference algorithms or defaulted genders which indicates no fairness benefits.

Average Exposure Ratio

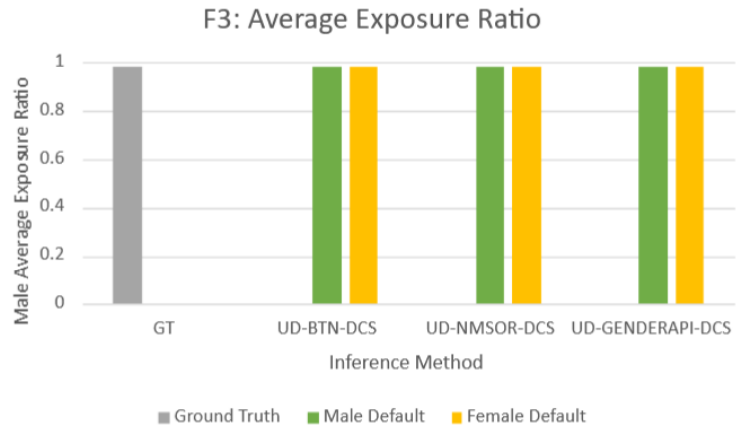


Figure 41: Average Exposure Ratio using UD-GT-DCS and UD-INF-DCS.

Figure 41 shows that the Average Exposure Ratios that are equal across all compared rankings. This value approaches one but shows that males are under-represented in the final rankings.

Average Positional Difference in Skew

Table 18 shows the average positional difference in skew of a particular group by comparing two different rankings. The first ranking is produced by the post-processing fair ranking algorithm with ground-truth information and the second is produced by the post-processing

fair ranking algorithm with inferred demographic information. A positive number means that group experienced an increase in the overall skew across all positions meaning that they became more represented when using inferred information. A negative number means that group experienced a decrease in the overall skew across all positions meaning that they became more under-represented when using inferred information. A value of 0 indicates that there was no change in representation between the rankings.

Ranking 1 => Ranking 2	**Males**	**Females**
UD-GT-DCS => UD-BTN-DCS (Male Default)	0.0	0.0
UD-GT-DCS => UD-BTN-DCS (Female Default)	0.0	0.0
UD-GT-DCS => UD-NameSor-DCS (Male Default)	0.0	0.0
UD-GT-DCS => UD-NameSor-DCS (Female Default)	0.0	0.0
UD-GT-DCS => UD-GAPI-DCS (Male Default)	0.0	0.0
UD-GT-DCS => UD-GAPI-DCS (Female Default)	0.0	0.0

Table 18: Average Positional Difference in Skew UD-GT-DCS and UD-INF-DCS.

The Average Positional Difference in Skew observed in the table above (Table 18) where all 0.0, meaning there was no change in position of either groups.

Conclusion.

DetConstSort re-rankings did not seem to provide any significant differences across inference algorithms when compared to the ground truth.

Utility Question 3

How does the utility of rankings produced by post-processing a ranking from a fairness-unaware LTR differ when using inferred demographics versus ground-truth demographics?

Average Positional Difference in NDCG

Table 19 shows the average positional difference in NDCG when ranking with DetConstSort in combination with ground-truth demographics versus inferred demographics. A negative value indicates that there was an average loss in utility when using inferred demographics whereas a positive value indicates that there was an average gain in utility when using inferred demographics and a 0 indicating no change.

Ranking Comparison	Average Positional Difference
UD-GT-DCS => UD-BTN-DCS (Male Default)	-0.023617040112609 94
UD-GT-DCS => UD-BTN-DCS (Female Default)	0.004291891095447243
UD-GT-DCS => UD-NameSor-DCS (Male Default)	-0.0010569823425945904
UD-GT-DCS => UD-NameSor-DCS (Female Default)	0.003884424184950575
UD-GT-DCS => UD-GAPI-DCS (Male Default)	-0.0006401800812588549
UD-GT-DCS => UD-GAPI-DCS (Female Default)	0.0011322997548662114

Table 19: Average Positional Difference in NDCG using UD-GT-DCS and UD-INF-DCS.

In Table 19 the Average Positional Difference in NDCG is relevant to the default gender values of the inference methods. The default female inference rankings increase the utility value when compared to the ground truth ranking. While when male is defaulted in the inference ranking, utility decreases in all three inference algorithms.

Fairness Question 4

How does the fairness of rankings obtained using ground-truth demographics differ from a fairness-aware in-processing LTR model to a post-processing fair ranking algorithm?

NDKL

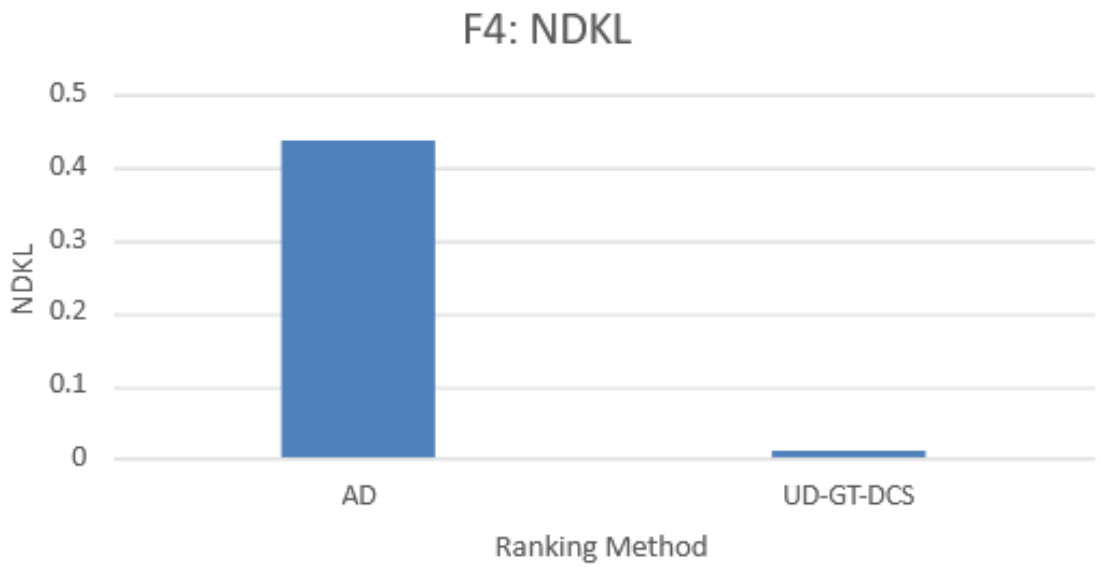


Figure 42: NDKL using AD-GT and UD-GT-DCS.

The results observed from Figure 42 show that DetConstSort produced a ranking much closer to zero than fairness-aware DELTR did on the ground truth dataset. This would mean that DetConstSort ranked more fairly under the criteria of NDKL than the fairness-aware DELTR model.

Average Exposure Ratio

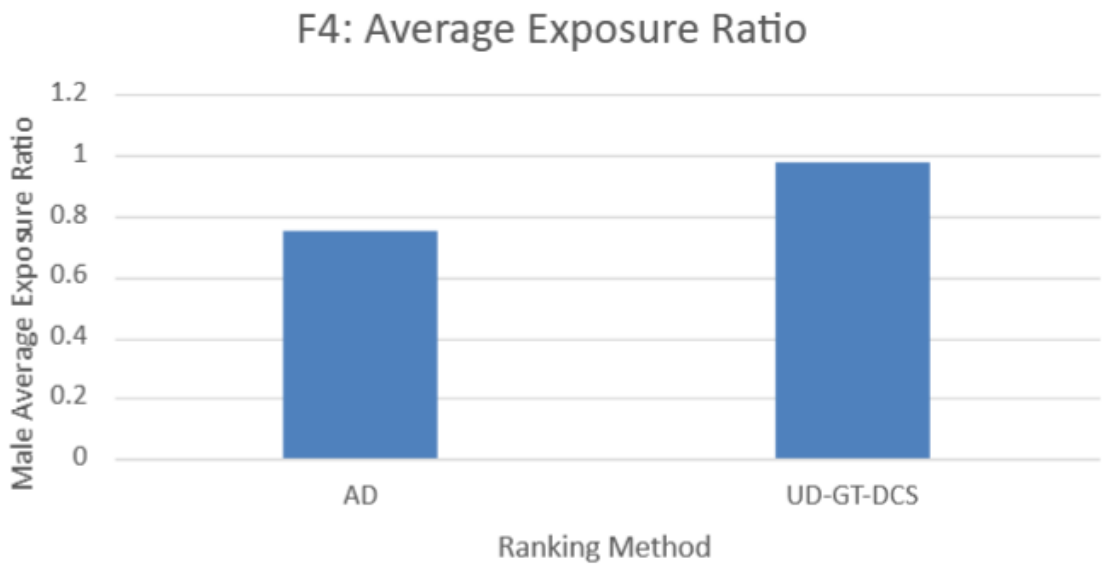


Figure 43: Average Exposure Ratio using AD-GT and UD-GT-DCS.

As seen in Figure 43 above, the AD-GT ranking has a Average Exposure Ratio of 0.748441379, while the DetConstSort ranking value is 0.980580594. Having a value closer to one is the ideal

for equal exposure. which means ranking produced by DetConstSort was more fair.

Average Positional Difference in Skew Table 20 shows the average difference in skew when ranking ground-truth demographic information in combination with fairness-aware DELTR versus DetConstSort. A negative value indicates that the specified group saw a decrease in representation when using DetConstSort whereas a positive value indicates that the specified group saw an increase in representation when using DetConstSort.

Ranking 1 ⇒ Ranking 2	**Males**	**Females**
AD-GT ⇒ UD-GT-DCS	0.6637150393271845	-0.6637150393271845

Table 20: Average Positional Difference in Skew using AD-GT and UD-GT-DCS.

Table 20 shows a negative change in the skew for the female group and a positive change in the male group. The females had a decrease in visibility in the DetConstSort ranking when compared to the DELTR ranking.

The results in fairness question 4 show that in the Boston Marathon experiment, under ground-truth demographic information, the post-processing fair ranking algorithm (DetConstSort) yielded more equally distributed based on all three fairness metrics.

Utility Question 4 *How does the utility of rankings obtained from a fairness-aware in-processing LTR model, ranked using ground-truth demographics, compare to the utility of rankings obtained from post-processing a ranking from a fairness-unaware LTR?*

Average Positional Difference in NDCG

Table 21 shows the average positional difference in NDCG when ranking ground-truth information in combination with fairness-aware DELTR versus DetConstSort. A positive value indicates that there was an increase in utility when ranking with DetConstSort using ground-truth demographic information whereas a negative value indicates a decrease in average utility.

Ranking Comparison	**Average Positional Difference**
AD-GT => UD-GT-DCS	-0.08426402858410362

Table 21: Average Positional Difference in NDCG using AD-GT and UD-GT-DCS.

Having a negative Average Positional Difference NDCG value shows an overall decrease in utility in the post-processing fairness ranking. The DELTR ranking's utility performed better than DetConstSort.

Fairness Question 5 *How does the fairness of rankings obtained from a fairness-aware in-processing LTR model, ranked using inferred demographics, compare to the fairness of rankings obtained from a post-processing fair ranking algorithm, ranked using inferred demographics?*

NDKL

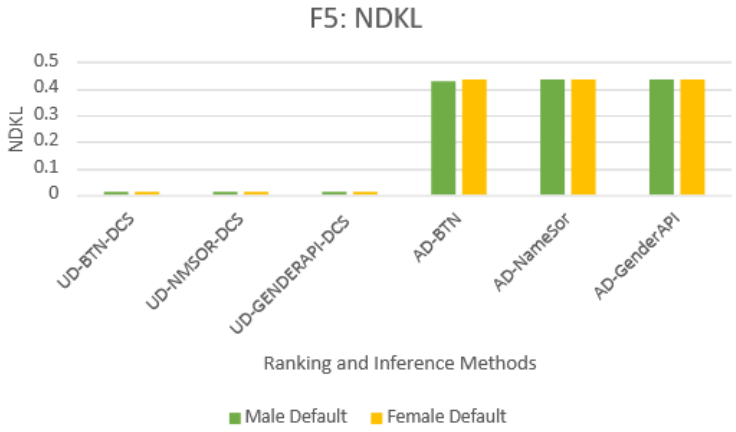


Figure 44: NDKL using AD-INF and UD-INF-DCS.

As seen in the table above (Figure 44), the post-processing ranking performs more fairly across all inference methods. The NDKL values for the DetConstSort rankings are much closer to zero than the DELTR fairness-aware inference rankings.

Average Exposure Ratio

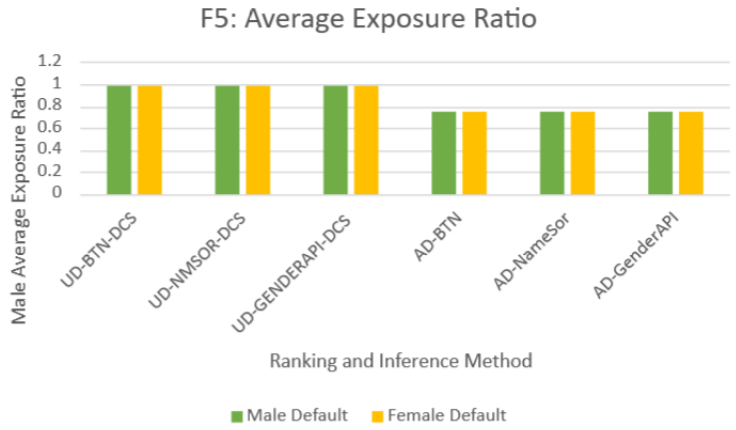


Figure 45: Average Exposure Ratio using AD-INF and UD-INF-DCS.

The Average Exposure Ratio give insight whether males are over or under-represented. In Figure 45, the values of the in-processing ranking are further from one, meaning the males are under-represented. While DetConstSort produces a more evenly represented ranking across all inference algorithms.

Average Positional Difference in Skew

Table 22 shows the average positional difference in skew when ranking inferred demographic information in combination with fairness-aware DELTR versus DetConstSort. A positive value indicates that the specified group saw an increase in average representation when using DetConstSort and a negative value indicates that the specified group saw a decrease in average representation when using DetConstSort.

Ranking 1 => Ranking 2	Males	Females
AD-BTN => UD-BTN-DCS (Male Default)	0.6517921763315834	-0.6517921763315833
AD-BTN => UD-BTN-DCS (Female Default)	0.6611575255156613	-0.6611575255156613
AD-NameSor => UD-NameSor-DCS (Male Default)	0.6618949751804253	-0.6618949751804253
AD-NameSor => UD-NameSor-DCS (Female Default)	0.6616556894642583	-0.6616556894642583
AD-GAPI => UD-GAPI-DCS (Male Default)	0.6621238175437689	-0.6621238175437689
AD-GAPI => UD-GAPI-DCS (Female Default)	0.6628742116344679	-0.6628742116344679

Table 22: Average Positional Difference in Skew using AD-INF and UD-INF-DCS.

The Average Positional Difference in Skew seen in Table 22 shows an increase in male exposure but a decrease in female exposure.

As seen in Fairness Question 4, DetConstSort ranks more fairly than DELTR fairness-aware models. The findings from Fairness Question 5 show that this is still true across all three inference methods as well.

Utility Question 5 *How does the utility of rankings obtained from a fairness-aware in-processing LTR model, ranked using inferred demographics, compare to the utility of rankings obtained from a post-processing fair ranking algorithm, ranked using inferred demographics?*

Average Positional Difference in NDCG

Table 23 shows the average positional difference in NDCG when ranking inferred demographic information in combination with fairness-aware DELTR versus DetConstSort. A negative value indicates that the ranking saw an average decrease in utility when using DetConstSort whereas a positive value indicates that the ranking saw an average increase in utility when using DetConstSort.

Ranking Comparison	Average Positional Difference
AD-BTN => UD-BTN-DCS (Male Default)	-0.06860352694975036
AD-BTN => UD-BTN-DCS (Female Default)	-0.08026461688165264
AD-NameSor => UD-NameSor-DCS (Male Default)	-0.08233923577239351
AD-NameSor => UD-NameSor-DCS (Female Default)	-0.08106595901658402
AD-GAPI => UD-GAPI-DCS (Male Default)	-0.08221906616383183
AD-GAPI => UD-GAPI-DCS (Female Default)	-0.08225545142161209

Table 23: Average Positional Difference in NDCG using AD-GT and UD-GT-DCS.

The utility consistently decreased across all inference methods in the DetConstSort ranking, as seen in Table 23. DELTR fairness-aware preforms better with the Average Positional Difference in NDCG and has a better utility score, regardless of which inference method is used.

4.3 Cherry Blossom Experiment Results

4.3.1 Cleaning and Splitting Cherry Blossom

The Cherry Blossom dataset was cleaned and ranked according to the score attribute which is the pace seconds. This means that the racers with the lowest pace time, in seconds, will be ranked highest and the racers with the highest pace time will be ranked the lowest.

For these experiments, we chose to keep the train-test split the same as the previous datasets. 80% of the data was used to train the models, and the remaining 20% of the data was used for testing the models.

The original Cherry Blossom dataset was 19,961 records long. In the dataset, approximately 38.5% were male and approximately 61.5% were female. In the initial experimental stages, we found that it was not feasible to use the entire large dataset due to restraints on time and computing power. As a result, we began strategically downsampling the dataset.

In the first attempt at downsampling, the goal was to have 10,000 records to train the models on. Therefore, the size of the downsampled dataset needed to be 12,500 to maintain the 80/20 split desired. We also wanted to maintain the proportion of males to females in the original dataset. This downsampled dataset consisted of 4,819 males and 7,682 females. However, this dataset was still too large for the restraints on time and computing power.

In the second round of down-sampling, the goal was to have the total dataset consist of 5,000 records. This number was chosen as the NBA/WNBA experiments had been successfully completed with that amount of records. This dataset consisted of 1,927 males and 3,073 females, and was chosen as the final downsampled dataset to be used for the Cherry Blossom experiments.

4.3.2 Training With Cherry Blossom Data

Multiple attempts at training the models on this data were required. For these experiments, it was necessary to consider the gamma values, the number of epochs to train the models on, and the amount of data to use. The gamma values chosen were 1.0 for fairness-aware, and 0.0 for fairness-unaware. These values were chosen as they had provided meaningful results in [6]. Originally, we had planned to use 10,000 epochs to train the models. However, when facing restrains on time and computing power, we decided to lower the number of epochs to 5,000.

4.3.3 Inferring Cherry Blossom Data

Using the 20% test split, we developed inferred datasets with inferred genders for the Cherry Blossom participants. We used the following inference algorithms:

- Behind The Name
- NameSor
- GenderAPI

When inferring the test data on each inference algorithm there was a male default inferred result and a female default inferred result. This is because, at times the inference algorithms were not able to produce a predicted gender from a name, in these cases we defaulted all the unknowns to male for one data set and all the unknowns to female in another data set. We describe the inference algorithm accuracy statistics in Table 24.

	Percent Unidentifiable	Accuracy Excluding Unidentifiable Results	Accuracy Including Unidentifiable Results	% Female in Unidentified	% Male in Unidentified	% Accuracy when default is Male	% Accuracy when default is Female
Behind the Name	15.1%	98.9%	83.9%	78.8%	21.2%	87.1%	95.8%
NameSor	5.8%	98.5%	92.7%	55.0%	45.0%	95.4%	95.9%
GenderAPI	0.68%	96.4%	95.7%	57.1%	42.9%	96.0%	96.1%

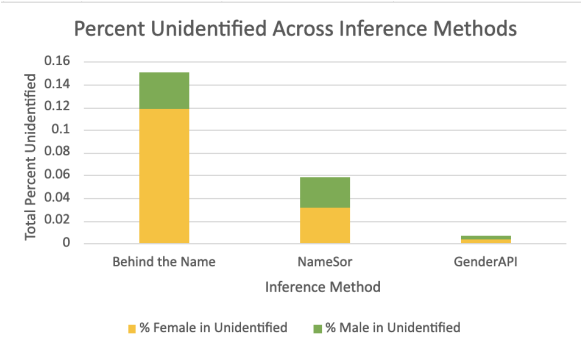
Table 24: Statistics found when inferring gender using Cherry Blossom Participant Names

Behind the Name. This inference algorithm was unable to infer the gender of 15.1% of names in the Cherry Blossom dataset. Of those unidentified names, 78.8% were ground-truth female and 21.2% were ground-truth male. When excluding the unidentifiable names, Behind the Name was 98.9% accurate with respect to the ground-truth gender of the names in the dataset. When including the unidentifiable names, Behind the Name was still 83.9% accurate with respect to the ground-truth gender of the names in the dataset. When the unidentified names were assigned male by default, the inference algorithm was only 87.1% accurate. When the unidentified names were assigned female by default, the inference algorithm was 95.8% accurate.

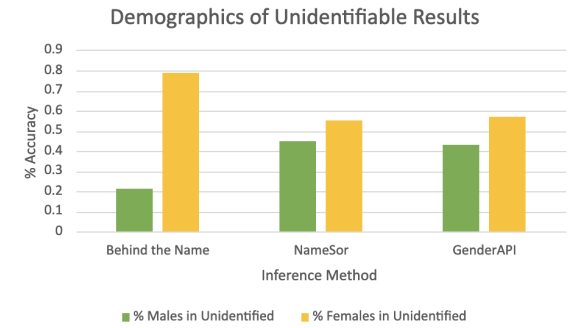
NameSor. This inference algorithm was unable to infer the gender of 5.8% of names in the Cherry Blossom dataset. Of those unidentified names, 55.0% were ground-truth female and 21.2% were ground-truth male. When excluding the unidentifiable names, NameSor was 98.5% accurate with respect to the ground-truth gender of the names in the dataset. When including the unidentifiable names, NameSor was only 92.7% accurate with respect to the ground-truth gender of the names in the dataset. When the unidentified names were assigned male by default, the inference algorithm was 95.4% accurate. When the unidentified names were assigned female by default, the inference algorithm was 95.9% accurate.

GenderAPI. This inference algorithm was the most accurate, as it was only unable to infer the gender of 0.68% of names in the Cherry Blossom dataset. Of those unidentified names, 57.1% were ground-truth female and 42.9% were ground-truth male. When excluding the unidentifiable names, GenderAPI was 96.4% accurate with respect to the ground-truth gender of the names in the dataset. When including the unidentifiable names, GenderAPI was only 95.7% accurate with respect to the ground-truth gender of the names in the dataset. When the unidentified names were assigned male by default, the inference algorithm was 96.0% accurate. When the unidentified names were assigned female by default, the inference algorithm was 96.1% accurate.

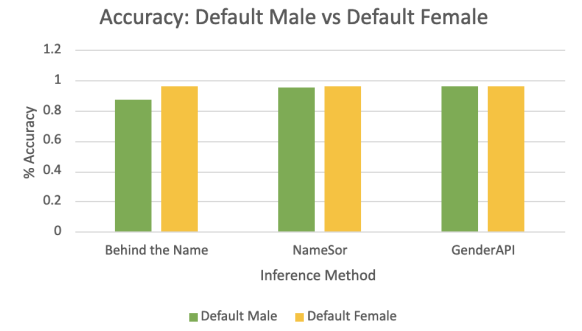
To summarize the accuracy statistics of the three inference methods, we can refer to the graphs shown in Figure 46.



(a) Percent Unidentified



(b) Demographics of Unidentified Names



(c) Accuracy with Different Default Values

Figure 46: Graphs describing inference accuracy statistics with Cherry Blossom data

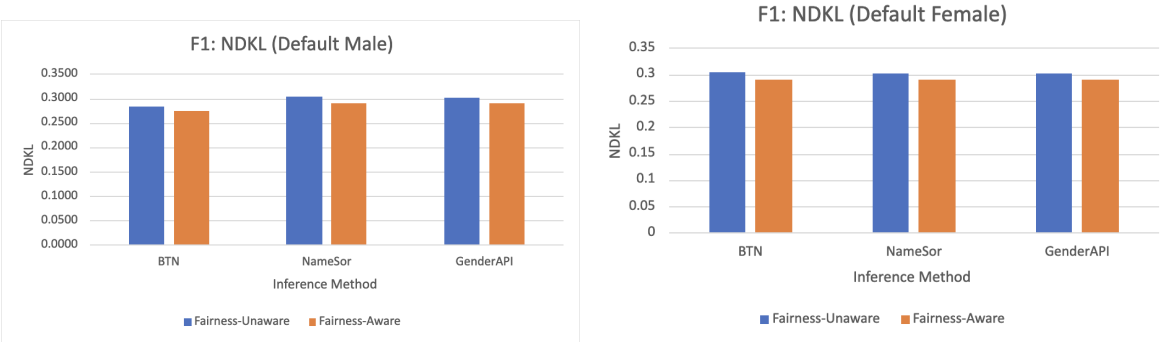
Figure 46a depicts the percent of names that were unidentified in the Cherry Blossom dataset by each inference algorithm. Behind the Name had the most unidentified results, and the majority of unidentifiable names were female for each inference algorithm. The demographics of the unidentified group of names are broken down in Figure 46b. Finally, in Figure 46c, we have the full picture of the varying levels of accuracy of each inference algorithm. Overall, the accuracy was consistently high for each algorithm when defaulting the gender to female. When defaulting to male, Behind the Name performed with less accuracy than the other two inference algorithms.

4.3.4 Research Questions

In this section we explore the different fairness and utility research questions described in the methodology with respect to the Cherry Blossom data set.

Fairness Question 1
Given uncertain demographic information, how does the fairness of rankings produced from a fairness-unaware LTR model compare to the fairness of rankings produced from a fairness-aware LTR model?

NDKL

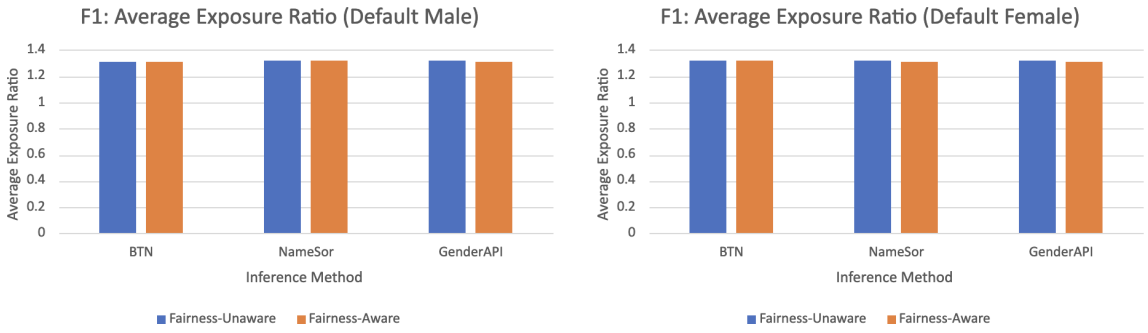


(a) NDKL Male Default (b) NDKL Female Default

Figure 47: F1, NDKL, Cherry Blossom

Overall, the NDKL values for all of the rankings related to this research question were very similar in value, as shown in the graphs from Figure 47. For each ranking, NDKL was lower when ranking with fairness-aware DELTR than when ranking with fairness-unaware DELTR. When considering the rankings where default values were male, the NDKL was consistently the lowest in rankings where the protected attribute of gender was inferred with Behind the Name. However, when the default values were female, the NDKL was lowest in rankings where the protected attribute of gender was inferred with GenderAPI.

Average Exposure Ratio



(a) Avg. Exp. Ratio: Male Default (b) Avg. Exp. Ratio: Female Default

Figure 48: F1, Average Exposure Ratio, Cherry Blossom

The average exposure ratio values, shown in Figure 48, were also very similar to one another for all rankings that were relevant to this research question. For each ranking, the average exposure ratio was closer to 1 when ranking with fairness-aware DELTR than when ranking with

fairness-unaware DELTR. When considering the rankings where the default gender was male, in Figure 48a, the average exposure ratio was closer to 1 in the rankings where the protected attribute of gender was inferred with Behind the Name. However, when the default gender was female, in Figure 48b, the average exposure ratio was nearly identical across inference algorithms; rankings where gender was inferred with NameSor and Gender API did slightly better in this instance when ranking with fairness-aware DELTR.

Average Positional Difference in Skew

In Table 25, we see the average positional difference values for the male and female skew when comparing rankings done with fairness-aware DELTR and fairness-unaware DELTR, where the gender was inferred.

Ranking 1 ⇒ Ranking 2	Males	Females
UD-BTN (Male Default) ⇒ AD-BTN (Male Default)	-0.00977081280472612	0.0058534126245031
UD-BTN (Female Default) ⇒ AD-BTN (Female Default)	-0.012600820480065333	0.0075487887396057
UD-NameSor (Male Default) ⇒ AD-NameSor (Male Default)	-0.011830838697975136	0.007087514823709556
UD-NameSor (Female Default) ⇒ AD-NameSor (Female Default)	-0.012274534308857969	0.007353320089052696
UD-GAPI (Male Default) ⇒ AD-GAPI (Male Default)	-0.012191884205578348	0.007303806791886721
UD-GAPI (Female Default) ⇒ AD-GAPI (Female Default)	-0.012265401363469342	0.00734784880443132

Table 25: F1, Avg. Positional Difference in Skew, Cherry Blossom

According to Table 25, across all inference algorithms, there was an increase in skew, or representation, for the female group when using fairness-aware DELTR. More specifically, there was a higher average increase in skew for the female group when the unknown demographics from the inference algorithms were defaulted to female. Finally, when the gender was inferred by Behind the Name, the skew of the rankings increased more than when inferring with other inference algorithms.

Conclusion.

In summary, the rankings were all more fair with respect to the chosen metrics when ranking with fairness-aware DELTR. When the default gender was male, rankings where gender was inferred by Behind the Name were often the most fair, of the three inference algorithms. Alternatively, when the default gender was female, rankings where gender was inferred by GenderAPI were the most fair, of the the three inference algorithms.

Utility Question 1

Given uncertain demographic information, how does the utility of rankings produced from a fairness-unaware LTR model compare to the utility of rankings produced from a fairness-aware LTR model?

Average Positional Difference in NDCG

Table 26 shows the average positional difference in the NDCG of two different rankings. The first ranking is produced by a fairness-unaware LTR model and the second is produced by a fairness-aware LTR model. A positive number means that the NDCG experienced an average increase across all positions meaning that the utility was increased when using a fairness-aware LTR model. A negative number means that the NDCG experienced an average decrease in all positions meaning that the ranking lost utility.

Ranking Comparison	Average Positional Difference
UD-BTN (Male Default) ⇒ AD-BTN (Male Default)	0.7308728506240818
UD-BTN (Female Default) ⇒ AD-BTN (Female Default)	0.7317818016762968
UD-NameSor (Male Default) ⇒ AD-NameSor (Male Default)	0.7312488866983278
UD-NameSor (Female Default) ⇒ AD-NameSor (Female Default)	0.7317074348634717
UD-GAPI (Male Default) ⇒ AD-GAPI (Male Default)	0.7314415043184441
UD-GAPI (Female Default) ⇒ AD-GAPI (Female Default)	0.7314886881899813

Table 26: U1, Avg. Positional Difference in NDCG, Cherry Blossom

In Table 26, we can observe how using fairness-aware DELTR produced rankings of higher utility across all inference algorithms. In particular, when the protected attribute of gender were defaulted to female in a ranking, fairness-aware DELTR was able to produce a ranking with slightly higher utility than when the default was male. Finally, we can also observe that the highest increase in NDCG occurred when using gender inferred by Behind the Name. Therefore, we can conclude that when using fairness-aware DELTR, we produce rankings with higher utility when inferring gender with less accuracy.

Fairness Question 2

How does the fairness of rankings produced from a fairness-aware LTR model compare when ranking with ground-truth demographic information and when ranking with inferred demographic information?

NDKL

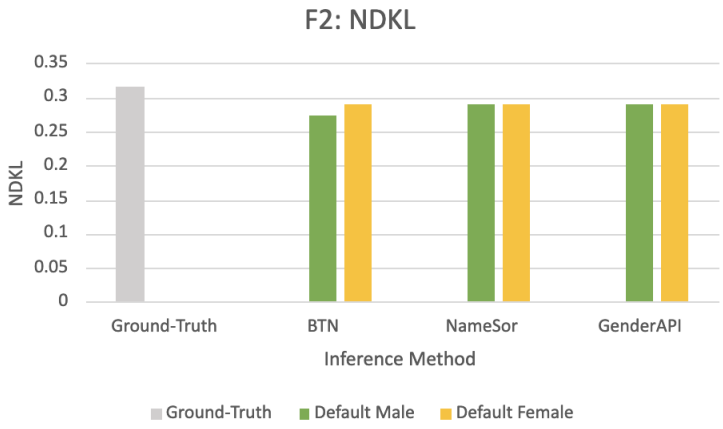


Figure 49: F2, NDKL, Cherry Blossom

For all rankings produced with inferred demographic information, the NDKL was closer to zero than the rankings produced with ground-truth demographic information (see Figure 49). All of the rankings produced with inferred gender had approximately the same NDKL value. The ranking with a notably lower NDKL was the ranking where gender was inferred with Behind the Name, and records with an unidentified gender were given a default value of male.

Average Exposure Ratio

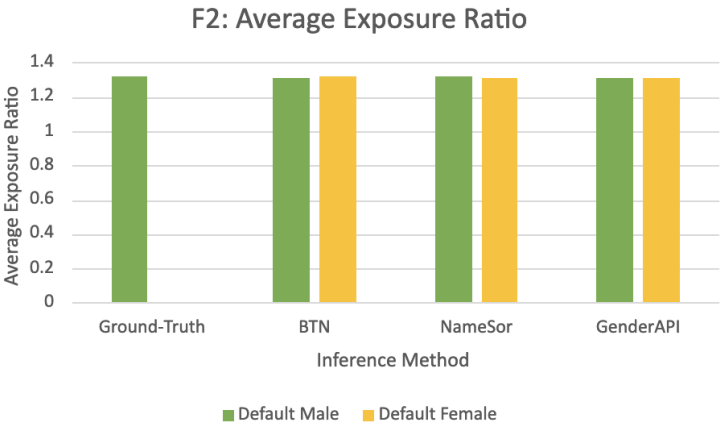


Figure 50: F2, Average Exposure Ratio, Cherry Blossom

As shown by Figure 50, rankings produced with inferred demographic information had average exposure ratio that were slightly closer to one than rankings produced with ground-truth demographic information. The inference algorithm used did not seem to make a significant impact on these results.

Average Positional Difference in Skew

In Table 27, we see the average positional difference values for the male and female skews when comparing fair rankings produced with ground-truth gender information and inferred gender information.

	Males	Females
AD-GT ⇒ AD-BTN (Male Default)	-0.03315412777789492	0.019861683359203314
AD-GT ⇒ AD-BTN (Female Default)	-0.013997348951992216	0.008385408737493782
AD-GT ⇒ AD-NameSor (Male Default)	-0.015026709907188024	0.009002069247804594
AD-GT ⇒ AD-NameSor (Female Default)	-0.014879893374571913	0.008914115690339536
AD-GT ⇒ AD-GAPI (Male Default)	-0.01451162699555743	0.008693497906007324
AD-GT ⇒ AD-GAPI (Female Default)	-0.014195965708318365	0.008504394317521995

Table 27: F2, Avg. Positional Difference in Skew, Cherry Blossom

We can see that the representation of the female group increases in all rankings produced with inferred demographic information. In particular, the ranking with the greatest increase in representation for the female group was the ranking with gender inferred by Behind the Name. However, the difference between male and female defaults was marginal.

Conclusion

In the Cherry Blossom experiments, the results for this research question show that using inference to produce gender information produces more fair rankings. Additionally, we continue to observe that using Behind the Name, the least accurate inference algorithm, produces the most fair rankings when used in conjunction with fairness-aware DELTR.

Utility Question 2

How does the utility of rankings produced from a fairness-aware LTR model compare when ranking with ground-truth demographic information and when ranking with inferred demographic information?

Average Positional Difference in NDCG

Table 28 shows the average positional difference in the NDCG of two different rankings. The first ranking is produced by a fairness-aware LTR model with ground-truth gender, and the second is produced by a fairness-aware LTR model with inferred gender. A positive number means that the NDCG experienced an average increase across all positions meaning that the utility was increased when using a fairness-aware LTR model. A negative number means that the NDCG experienced an average decrease in all positions meaning that the ranking lost utility.

Ranking Comparison	Average Positional Difference
AD-GT ⇒ AD-BTN (Male Default)	0.004760794997140874
AD-GT ⇒ AD-BTN (Female Default)	-0.003364750671478449
AD-GT ⇒ AD-NameSor (Male Default)	0.0018073426010417996
AD-GT ⇒ AD-NameSor (Female Default)	-0.0024805949189740033
AD-GT ⇒ AD-GAPI (Male Default)	9.295162599082946e-05
AD-GT ⇒ AD-GAPI (Female Default)	-0.000385789466644441

Table 28: U2, Avg. Positional Difference in NDCG, Cherry Blossom

In Table 28, we can observe how using fairness-aware DELTR produced rankings of higher utility across all inference algorithms. In particular, when the protected attribute of gender was defaulted to male in a ranking, fairness-aware DELTR was able to produce a ranking with higher utility than when the default was female. Finally, we can also observe that the highest increase in NDCG occurred when using gender inferred by Behind the Name. Therefore, we continue to observe that rankings produced with gender inferred by Behind the Name have more utility than rankings produced with ground-truth gender.

Fairness Question 3

How does the fairness of rankings produced by post-processing a ranking from a fairness-unaware LTR differ when using inferred demographics versus ground-truth demographics?

NDKL

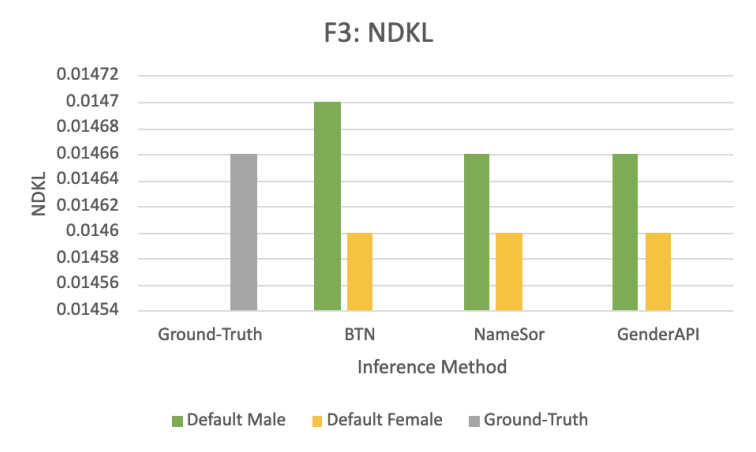


Figure 51: F3, NDKL, Cherry Blossom

As seen in Figure 51, the rankings produced almost identical NDKL values[3]. For all rankings produced with inferred demographic information, where the default value for gender was female, the NDKL was closer to zero than the ranking produced with ground-truth gender. However, when inferring and the default value for gender was male, the ranking produced with ground-truth demographic information had an NDKL value closer to zero. When the default value was female, each ranking with inferred gender had almost identical NDKL values.

Average Exposure Ratio

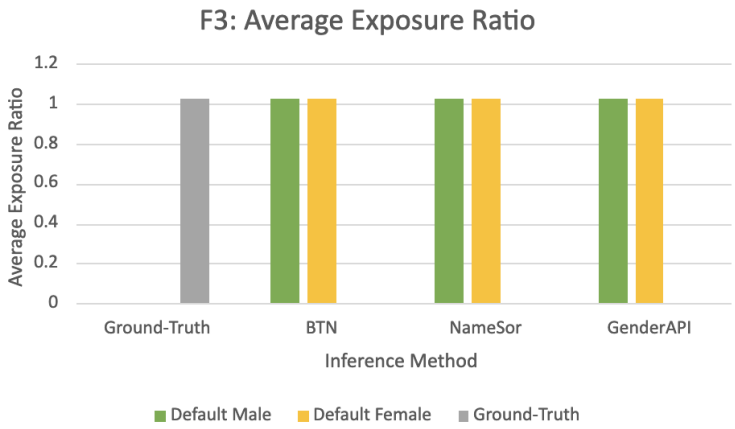


Figure 52: F3, Average Exposure Ratio, Cherry Blossom

As shown by Figure 52, all of the rankings that were relevant to this research question had the same average exposure ratio, to the seventh decimal place.

[3]Note that the y-axis begins at 0.01454 in order to better visualize the (otherwise minute) differences in NDKL values.

Average Positional Difference in Skew

In Table 29, we see the average positional difference values for the male and female skews when comparing fair rankings produced with ground-truth gender information and inferred gender information.

	Males	Females
UD-GT-DCS ⇒ UD-BTN-DCS (Male Default)	0.0	0.0
UD-GT-DCS ⇒ UD-BTN-DCS (Female Default)	0.0	0.0
UD-GT-DCS ⇒ UD-NameSor-DCS (Male Default)	0.0	0.0
UD-GT-DCS ⇒ UD-NameSor-DCS (Female Default)	0.0	0.0
UD-GT-DCS ⇒ UD-GAPI-DCS (Male Default)	0.0	0.0
UD-GT-DCS ⇒ UD-GAPI-DCS (Female Default)	0.0	0.0

Table 29: F3, Avg. Positional Difference in Skew, Cherry Blossom

Interestingly, each of the average positional difference values in Table 29 are 0.0. It is important to note that DetConstSort is a re-ranking algorithm, so it is more deterministic than DELTR, an in-processing learning-to-rank model.

Conclusion.

In the Cherry Blossom experiment, the use of inferred gender versus ground-truth gender did not significantly impact the fairness of the rankings produced by DetConstSort. Interestingly, when the rankings had a default gender of female, fairness metric NDKL showed that the rankings were more fair than when using ground-truth gender.

Utility Question 3

How does the utility of rankings produced by post-processing a ranking from a fairness-unaware LTR differ when using inferred demographics versus ground-truth demographics?

Average Positional Difference in NDCG

Table 30 shows the average positional difference in the NDCG of two different rankings. The first ranking is produced by DetConstSort with ground-truth gender, and the second is produced by DetConstSort with inferred gender. A positive number means that the NDCG experienced an average increase across all positions meaning that the utility was increased when using a fairness-aware LTR model. A negative number means that the NDCG experienced an average decrease in all positions meaning that the ranking lost utility.

Ranking 1 ⇒ Ranking 2	Average Positional Difference in NDCG
UD-GT-DCS ⇒ UD-BTN-DCS (Male Default)	0.011303426392423094
UD-GT-DCS ⇒ UD-BTN-DCS (Female Default)	-0.0014141577647209123
UD-GT-DCS ⇒ UD-NameSor-DCS (Male Default)	0.004387519899326481
UD-GT-DCS ⇒ UD-NameSor-DCS (Female Default)	-0.00041300399944901476
UD-GT-DCS ⇒ UD-GAPI-DCS (Male Default)	0.0024252240311912116
UD-GT-DCS ⇒ UD-GAPI-DCS (Female Default)	0.0018751789030113358

Table 30: U3, Avg. Positional Difference in NDCG, Cherry Blossom

In Table 30, we can observe how rankings produced with inferred gender do not always have a higher utility than the rankings produced with ground-truth gender data. Across all rankings relevant to this research question, when the default value for gender was male, the average positional difference in NDCG was positive. However, when the default value was female, the average positional difference in NDCG was negative. In particular, when inferring gender with Behind the Name, defaulting to male when it could not infer a person's gender, we observe the highest average positional difference in NDCG. Since Behind the Name was the least accurate inference algorithm for the Cherry Blossom dataset when defaulting to male, we can conclude that DetConstSort can produce a ranking with higher utility when using a less accurate inference algorithm.

Fairness Question 4

How does the fairness of rankings obtained using ground-truth demographics differ from a fairness-aware in-processing LTR model to a post-processing fair ranking algorithm?

NDKL

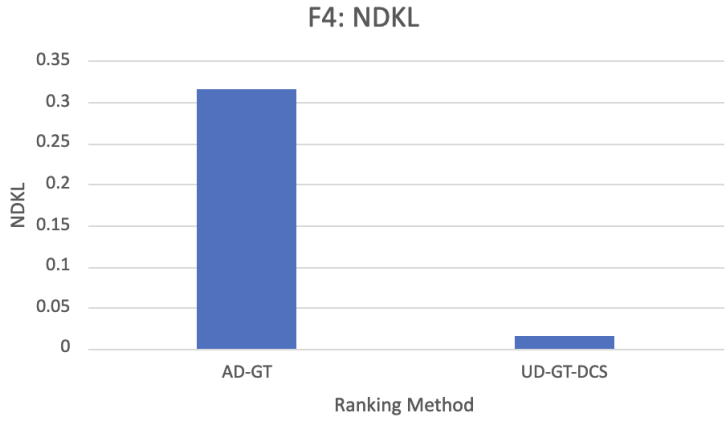


Figure 53: F4, NDKL, Cherry Blossom

Figure 53 shows the difference in NDKL values for rankings produced by fairness-aware DELTR and by DetConstSort, both with ground-truth gender information. The ranking produced by DetConstSort has an NDKL value of approximately 0.0147. This is much closer to the ideal value of zero than the NDKL for the ranking produced by fairness-aware DELTR, which is approximately 0.3155.

Average Exposure Ratio

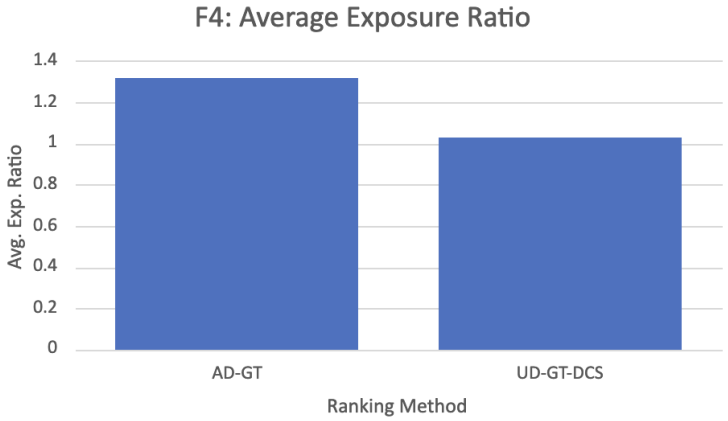


Figure 54: F4, Average Exposure Ratio, Cherry Blossom

Figure 54 depicts the difference in the average exposure ratio for rankings produced by fairness-aware DELTR and by DetConstSort, both with ground-truth gender information. The ranking produced by DetConstSort has a average exposure ratio of approximately 1.0281. This is much closer to the ideal value of one than the ratio for the ranking produced by fairness-aware DELTR, which is approximately 1.3214.

Average Positional Difference in Skew

In Table 31, we see the average positional difference values for the male and female skews when comparing fair rankings produced with fairness-aware DELTR and DetConstSort.

	Males	Females
AD-GT ⇒ UD-GT-DCS	-0.7606501717430542	0.4556836168182093

Table 31: F4, Avg. Positional Difference in Skew, Cherry Blossom

As seen in Table 31, the average positional difference in skew between these two rankings is negative for males and positive for females. Therefore, on average, the female group saw an increase in representation earlier in the ranking when using DetConstSort.

Conclusion

Overall, in the Cherry Blossom experiment, using DetConstSort with ground-truth gender produced more fair rankings than when using fairness-aware DELTR.

Utility Question 4 *How does the utility of rankings obtained from a fairness-aware in-processing LTR model, ranked using ground-truth demographics, compare to the utility of rankings obtained from post-processing a ranking from a fairness-unaware LTR?*

Average Positional Difference in NDCG

Table 32 shows the average positional difference in the NDCG of two different rankings. The first ranking is produced by fairness-aware DELTR with ground-truth gender, and the second is produced by DetConstSort with ground-truth gender. A positive number means that the NDCG experienced an average increase across all positions meaning that the utility was increased when using a fairness-aware LTR model. A negative number means that the NDCG experienced an average decrease in all positions meaning that the ranking lost utility.

Ranking Comparison	Average Positional Difference
AD-GT ⇒ UD-GT-DCS	-0.7729433802361542

Table 32: U4, Avg. Positional Difference in NDCG, Cherry Blossom

It is evident from Table 32 that the average positional difference of NDCG between these two rankings is negative. Therefore, we can conclude that using DetConstSort to rank data with ground-truth gender information produces a ranking with decreased utility than when using fairness-aware DELTR.

Fairness Question 5 *How does the fairness of rankings obtained from a fairness-aware in-processing LTR model, ranked using inferred demographics, compare to the fairness of rankings obtained from a post-processing fair ranking algorithm, ranked using inferred demographics?*

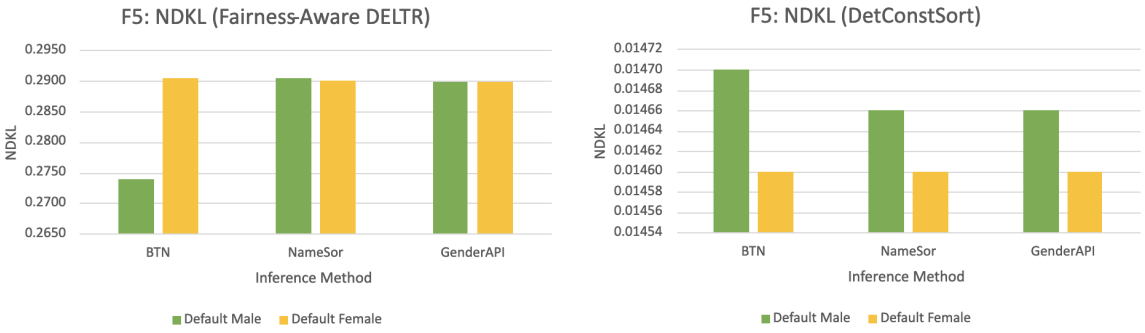
NDKL

(a) NDKL Male Default

(b) NDKL Female Default

Figure 55: F5, NDKL, Cherry Blossom: Difference in Default Gender

Figure 55a shows the difference in NDKL between fairness-aware DELTR and DetConstSort when the default gender is male. Figure 55b shows the difference between the two models when the default gender is female. It is evident from both of these figures that rankings using DetConstSort have an NDKL that is much closer to zero, the ideal value, than the rankings that use DELTR. For a clearer view of the change in NDKL between rankings, Figure 56 displays the data on a smaller y-axis, and splits the data in two graphs according to the ranking method used. [4]



(a) NDKL from DELTR

(b) NDKL from DetConstSort

Figure 56: F5, NDKL, Cherry Blossom: Difference between Ranking Methods

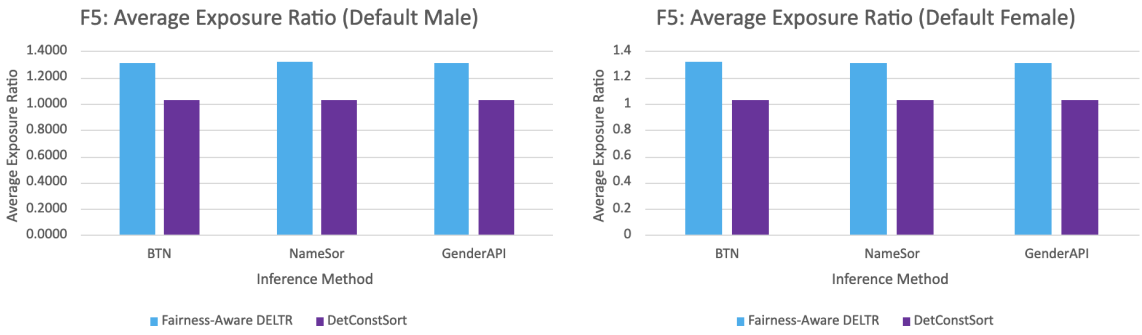
Figure 56a shows how the NDKL changes when the inference method changes, when using fairness-aware DELTR to produce a ranking. In this figure, we can observe that the NDKL only slightly changes with different inference methods. The main outlier is the ranking produced when gender was inferred with Behind the Name. That ranking has an NDKL of 0.2793.

Figure 56b shows how the NDKL changes when the inference method changes, when using DetConstSort to produce a ranking. We observe more of a disparity in NDKL in this case.

[4]Note that the y-axis on both subfigures are greater than zero, and are not equal to one another. This was done intentionally in order to display the change in NDKL between rankings on a small scale.

When the default gender was female, the NDKL stayed the same for all three inference methods at a value of 0.0146. However, when the default gender was male, the NDKL was slightly closer to zero when the gender was inferred by NameSor and Gender API. It is important to note that the changes in NDKL across these rankings are all less than 0.0003.

Average Exposure Ratio



(a) Default Male (b) Default Female

Figure 57: F5, Average Exposure Ratio, Cherry Blossom

Figure 57 depicts the average exposure ratio for the rankings relevant to this research question. While there are slight variances in the average exposure ratio of rankings produced by fairness-aware DELTR, the ratio remains constant for the rankings produced by DetConstSort at 1.028075970. To more closely consider the differences in the rankings produced by DELTR, see Figure 58. [5]

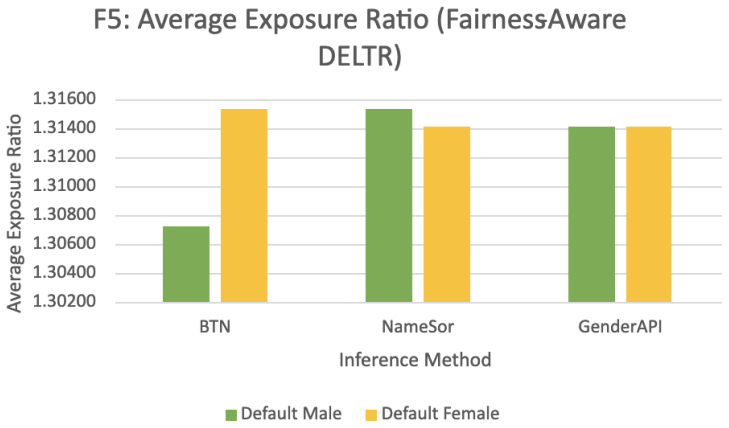


Figure 58: F5, Average Exposure Ratio, Cherry Blossom: DELTR

Figure 58 shows that while most of the rankings have a similar average exposure ratio, the notable exception is the ranking where gender was inferred with Behind the Name, and the default value was male. This ranking had an average exposure ratio of 1.3072.

Average Positional Difference in Skew

In Table 33, we see the average positional difference values for the male and female skews when comparing fair rankings produced with fairness-aware DELTR and DetConstSort, both with inferred gender information.

[5]Note that the lower bound of the y-axis is 1.3020 in order to view the difference in values more clearly.

Ranking 1 ⇒ Ranking 2	Males	Females
AD-BTN (Default Male) ⇒ UD-BTN-DCS (Default Male)	-0.7274960439651608	0.4358219334590057
AD-BTN (Default Female) ⇒ UD-BTN-DCS (Default Female)	-0.7466528227910633	0.44729820808071513
AD-NameSor (Default Male) ⇒ UD-NameSor-DCS (Default Male)	-0.7456234618358663	0.4466815475704047
AD-NameSor (Default Female) ⇒ UD-NameSor-DCS (Default Female)	-0.7457702783684831	0.4467695011278692
AD-GAPI (Default Male) ⇒ UD-GAPI-DCS (Default Male)	-0.7461385447474972	0.44699011891220186
AD-GAPI (Default Female) ⇒ UD-GAPI-DCS (Default Female)	-0.7464542060347359	0.4471792225006871

Table 33: F5, Avg. Positional Difference in Skew, Cherry Blossom

Table 33 shows that in general, the average positional difference in skew for the female group was positive, meaning their representation decreased when ranking with DetConstSort compared to fairness-aware DELTR. Therefore, the rankings produced with DetConstSort increased the representation of the under-represented group, females. The ranking where gender was inferred by Behind the Name, and the default value was female, saw the largest increase in representation for females from fairness-aware DELTR to DetConstSort.

In the Cherry Blossom experiment, DetConstSort produced significantly more fair rankings than fairness-aware DELTR when using inferred gender information. Although defaulting male or female did not make a difference when ranking with fairness-aware DELTR, defaulting female produced more fair rankings with DetConstSort.

Utility Question 5 *How does the utility of rankings obtained from a fairness-aware in-processing LTR model, ranked using inferred demographics, compare to the utility of rankings obtained from a post-processing fair ranking algorithm, ranked using inferred demographics?*

Average Positional Difference in NDCG

Table 34 shows the average positional difference in the NDCG of rankings produced with

fairness-aware DELTR and with DetConstSort, both with inferred demographic information.

Ranking 1 ⇒ Ranking 2	Average Positional Difference in NDCG
AD-BTN (Default Male) ⇒ UD-BTN-DCS (Default Male)	-0.7664007488408726
AD-BTN (Default Female) ⇒ UD-BTN-DCS (Default Female)	-0.7709927873293967
AD-NameSor (Default Male) ⇒ UD-NameSor-DCS (Default Male)	-0.7703632029378705
AD-NameSor (Default Female) ⇒ UD-NameSor-DCS (Default Female)	-0.7708757893166297
AD-GAPI (Default Male) ⇒ UD-GAPI-DCS (Default Male)	-0.7706111078309535
AD-GAPI (Default Female) ⇒ UD-GAPI-DCS (Default Female)	-0.7706824118664981

Table 34: U5, Avg. Positional Difference in NDCG, Cherry Blossom

It is evident from the table that all rankings produced with DetConstSort saw a decrease in utility than those produced with fairness-aware DELTR. However, the ranking that saw the lowest decrease in utility was when the gender was inferred by Behind the Name, with male as the default value for unidentified gender.

5 Conclusion

In this section, we summarize and discuss the final results of the experiment, contributions to the project, and potential future steps to continue this work.

5.1 Experiment Conclusions

When evaluating each research question across all three experiments, we consider the impact of the moving parts in the experiments: the various ranking methods and inference strategies.

Fairness Question 1
Given uncertain demographic information, how does the fairness of rankings produced from a fairness-unaware LTR model compare to the fairness of rankings produced from a fairness-aware LTR model?

In order to summarize this research question across all three experiments, and our chosen metrics, we consider the impact of a fairness-aware versus fairness-unaware DELTR, the impact of using the various inference algorithms, and the impact of defaulting the unknowns to male or female.

Across all three experiments, fairness-aware DELTR produced the most fair rankings when using inferred gender information. In the Boston Marathon Experiment, while some of the fairness metrics were consistent with the WNBA/NBA and Cherry Blossom experiments, the change in metrics was too small to make significant conclusions. The results from the WNBA/NBA and the Cherry Blossom experiments showed that the rankings were more fair when the gender was inferred by Behind the Name, and the default gender was male. Notably, the method of inferring gender with Behind the Name and defaulting gender to male is the least accurate with respect to the protected group, females.

Using these results, we can therefore conclude that when using a method of inference that is less accurate with respect to the protected group, we can produce more fair rankings with fairness-aware DELTR.

Utility Question 1
Given uncertain demographic information, how does the utility of rankings produced from a fairness-unaware LTR model compare to the utility of rankings produced from a fairness-aware LTR model?

In the WNBA/NBA experiment, when using fairness-aware DELTR and inferred gender information, we produce rankings with lower utility than when using fairness-unaware DELTR. In the Boston Marathon experiment, the change in utility was small enough to be negligible. However, in the Cherry Blossom experiment, using fairness-aware DELTR produced rankings with higher utility than when using fairness-unaware DELTR. When considering the ratio of protected group to non-protected group in each dataset, we observe that when the proportion of the protected group increases in the overall dataset, the utility of the ranking produced by fairness-aware DELTR, with inferred demographic information, increases.

It is also important to note that for all three experiments, the rankings with the highest utility were produced when using inference algorithms with lower accuracy.

Fairness Question 2
How does the fairness of rankings produced from a fairness-aware LTR model compare when ranking with ground-truth demographic information and when ranking with inferred demographic information?

In order to summarize this research question across all three experiments, and our chosen metrics, we will consider the impact of using ground-truth versus inferred demographic information under fairness-aware DELTR, the impact of using the various inference algorithms, and the impact of defaulting the unknowns to male or female.

Across all three experiments, fairness-aware DELTR combined with inferred demographic information produces equally or more fair rankings over using ground-truth demographic information. The most fair rankings occurred when inferring gender with Behind the Name, and when defaulting unidentifiable gender to male. We propose that using this inference method may have produced the most fair rankings because it minimized the perceived number of people in the protected group (female) for fairness-aware DELTR in the experiments. As a result, fairness-aware DELTR was given information that suggested less females were in the ranking, and therefore pushed them higher in the ranking to achieve a more fair result.

Utility Question 2
How does the utility of rankings produced from a fairness-aware LTR model compare when ranking with ground-truth demographic information and when ranking with inferred demographic information?

Across all three experiments, the changes in utility were small enough to be negligible. There were no discernable patterns in the increases or decreases of utility. Therefore, based on our results, we can conclude that using inference does not impact the utility of a ranking produced by fairness-aware DELTR.

Fairness Question 3
How does the fairness of rankings produced by post-processing a ranking from a fairness-unaware LTR differ when using inferred demographics versus ground-truth demographics?

In order to summarize this research question across all three experiments, and our chosen metrics, we will consider the impact of using ground-truth versus inferred demographic information under the post-processing fair ranking algorithm DetConstSort, the impact of using the various inference algorithms, and the impact of defaulting the unknowns to male or female.

Across all three experiments, the use of inferred gender versus ground-truth gender did not significantly impact the fairness of the rankings produced by DetConstSort. Interestingly, for the Cherry Blossom and WNBA/NBA experiments, when the rankings with inference had a default gender of female, the NDKL showed that the rankings were more fair than when using ground-truth gender. For the WNBA/NBA experiment, this conclusion was further justified by the skew values. It is also important to note that all three experiments had zero or close to zero average positional difference in skew. However, since the change in values when using inference versus ground-truth was so minor for all three experiments, we conclude that using inference with DetConstSort did not make an impact on the fairness of rankings.

Utility Question 3
How does the utility of rankings produced by post-processing a ranking from a fairness-unaware LTR differ when using inferred demographics versus ground-truth demographics?

According to the results of all three experiments, we can conclude that the change in utility of rankings when using inferred demographic information over ground-truth demographic information in combination with DetConstSort is negligible. It is important to note that in the WNBA/NBA and Cherry Blossom experiments, when the unknowns were defaulted to male across all inference algorithms, the utility of the rankings decreased. When the unknowns were defaulted to female across all inference algorithms, the utility of the rankings increased. Alternatively, the Boston Marathon experiment produced the opposite results for utility.

Fairness Question 4
How does the fairness of rankings obtained using ground-truth demographics differ from a fairness-aware in-processing LTR model to a post-processing fair ranking algorithm?

Across all three experiments and when using ground-truth demographic information, DetConstSort far outperformed fairness-aware DELTR when considering all three fairness metrics. We propose that this is due to DetConstSort being a deterministic algorithm that can be fine-tuned to achieve the desired results, whereas DELTR is a learning-to-rank model that can output unreliable rankings based on how the model was trained.

Utility Question 4 *How does the utility of rankings obtained from a fairness-aware in-processing LTR model, ranked using ground-truth demographics, compare to the utility of rankings obtained from post-processing a ranking from a fairness-unaware LTR?*

The change in utility when ranking ground-truth demographic information with DetConstSort over fairness-aware DELTR across all three experiments saw no discernible patterns or conclusions. In the WNBA/NBA experiment, utility increased by 0.32 when using DetConstSort over fairness-aware DELTR. However, in the Boston Marathon and Cherry Blossom experiments, utility decreased by 0.08 and 0.77 respectively. It is interesting to note that the proportion of the protected group (females) to the non-protected group (males) increased from WNBA/NBA to Boston Marathon to Cherry Blossom while the utility decreased.

Fairness Question 5 *How does the fairness of rankings obtained from a fairness-aware in-processing LTR model, ranked using inferred demographics, compare to the fairness of rankings obtained from a post-processing fair ranking algorithm, ranked using inferred demographics?*

Across all three experiments, DetConstSort produced significantly more fair rankings than fairness-aware DELTR when using inferred gender information. This remains true across all inference methods, whether we defaulted the unknowns to male or to female.

Utility Question 5 *How does the utility of rankings obtained from a fairness-aware in-processing LTR model, ranked using inferred demographics, compare to the utility of rankings obtained from a post-processing fair ranking algorithm, ranked using inferred demographics?*

The results to this question were very similar to Utility Question 4. The change in utility when ranking inferred demographic information with DetConstSort over fairness-aware DELTR across all three experiments again saw no discernible patterns. In the WNBA/NBA experiment, utility increased on average of 0.32 across all inference algorithms and male or female defaults.

However, for the Boston Marathon experiment, utility decreased on average of 0.07. Finally, for the Cherry Blossom Experiment, utility decreased on average of 0.77. Again, it is important to note that the proportion of the protected group to the non-protected group increased from WNBA/NBA to Boston Marathon to Cherry Blossom while the utility decreased.

5.2 Discussions

We study the effect of using inferred gender information in conjunction with different ranking methods on the fairness of a ranking. We found that we produced more fair rankings when using DetConstSort over DELTR. This could be due to the deterministic nature of DetConstSort as a post-processing re-ranking algorithm. DetConstSort allows for fine-tuning the desired characteristics of the ranking outputs, whereas rankings produced by DELTR, a learning-to-rank model, are far less reliable as they are heavily dependent on the methods of training and potential bias present in the dataset.

Additionally, we find that using inference methods that are less accurate with respect to the protected group produced more fair rankings. This conclusion was consistent when using DELTR as well as DetConstSort. However, these results were unexpected, as the previous work completed by [6] came to the opposite conclusions. Both our experiments and the experiments of [6] were limited by time. Due to the time constraints, we were not able to fine tune the gamma parameter of DELTR. Having a poorly fitted model could have had a significant effect on how our datasets were ranked. For future work, experimenting to find the optimal gamma value and number of iterations used when training the models will be critical for each individual dataset used. This will help to create models that are not over or under fitted to the data, allowing for more significant results.

We would also recommend that future work incorporate experiments on other protected attributes such as race, religion, age, etc. This would involve adjusting the FairRank package to accommodate such experiments, as well as conducting additional experiments on how inferring those protected attributes will affect the fairness of rankings.

5.3 Contributions

Marie Tessier, Sai Vadlamudi, and Brinda Venkataraman all contributed equally to the work on this project. Sai led the WNBA/NBA experiment and the development of the FairRank package. Brinda led the Cherry Blossom experiment, focused on the incorporation of fairness metrics, managing communications for the team, as well as contributing significantly to the FairRank package. Gabi led the Boston Marathon experiment, created the FairRank package structure and diagrams, and contributed significantly to the FairRank package.

5.4 Authorship

Section	Primary Writer	Editor
Abstract	Sai Vadlamudi Marie Tessier Brinda Venkataraman	Sai Vadlamudi
Executive Summary	Sai Vadlamudi Marie Tessier Brinda Venkataraman	Marie Tessier
Introduction	Sai Vadlamudi Marie Tessier Brinda Venkataraman	Brinda Venkataraman
Background	Sai Vadlamudi Marie Tessier Brinda Venkataraman	Sai Vadlamudi
Methods	Sai Vadlamudi Marie Tessier Brinda Venkataraman	Marie Tessier
WNBA/NBA Experiments and Analysis	Sai Vadlamudi	Brinda Venkataraman
Boston Marathon Experiments and Analysis	Marie Tesiier	Sai Vadlamudi
Cherry Blossom Experiments and Analysis	Brinda Venkataraman	Marie Tessier
Conclusions	Sai Vadlamudi Marie Tessier Brinda Venkataraman	Sai Vadlamudi Marie Tessier Brinda Venkataraman

References

[1] A. Ghosh, R. Dutt, and C. Wilson, "When Fair Ranking Meets Uncertain Inferences," May 2021.

[2] S. C. Geyik, S. Ambler, and K. Kenthapadi, "Fainess-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search," July 2019.

[3] A. Singh and T. Joachims, "Fairness of Exposure in Rankings," *Association for Computing Machinery*, Oct. 2018.

[4] A. Singh and T. Joachims, "Policy Learning for Fairness in Ranking," Feb. 2019.

[5] M. Zehlike and C. Castillo, "Reducing disparate exposure in ranking: A learning to rank approach," *Proceedings of The Web Conference 2020*, 2020.

[6] A. Pietrick, A. Romportl, S. Smith, O. Olulana, K. Cachel, and E. Rundensteiner, "Are Fair Learning To Rank Models Really Fair? An Analysis Using Inferred Gender," in *2022 IEEE MIT Undergraduate Research Technology Conference (URTC)*, pp. 1–5, Sept. 2022.

[7] "Equal Credit Opportunity Act (Regulation B) Ethnicity and Race Information Collection," Oct. 2017.

[8] I. Žliobaitė, "A survey on measuring indirect discrimination in machine learning," *ArXiv*, vol. abs/1511.00148, 2015.

[9] A. Mullick, S. Ghosh, R. Dutt, A. Ghosh, and A. Chakraborty, "Public sphere 2.0: Targeted commenting in online news media," in *European Conference on Information Retrieval*, pp. 180–187, Springer, 2019.

[10] Z. Ovaisi, R. Ahsan, Y. Zhang, K. Vasilaky, and E. Zheleva, "Correcting for selection bias in learning-to-rank systems," in *Proceedings of The Web Conference 2020*, pp. 1863–1873, 2020.

[11] A. Singh, N. Thakur, and A. Sharma, "A review of supervised machine learning algorithms," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 1310–1315, Ieee, 2016.

[12] A. Bower, L. Niss, Y. Sun, and A. Vargo, "Debiasing representations by removing unwanted variation due to protected attributes," *CoRR*, vol. abs/1807.00461, 2018.

[13] "Using publicly available information to proxy for unidentified race and ethnicity."

[14] "Baby Names from Social Security Card Applications - National Data." Type: dataset.

[15] M. Campbell, "The Meaning and History of First Names - Behind the Name."

[16] "Namsor: name checker for gender, origin and ethnicity classification."

[17] "Gender API About."