An External Fairness Evaluation of LinkedIn Talent Search

Tina Behzad[1], **Siddartha Devic**[2], **Vatsal Sharan**[2*], **Aleksandra Korolova**[3*], **David Kempe**[2*]

[1] Stony Brook University
[2] University of Southern California
[3] Princeton University
tbehzad@cs.stonybrook.edu, {devic, vsharan}@usc.edu, korolova@princeton.edu, David.M.Kempe@gmail.com

Abstract

We conduct an independent, third-party audit for bias of LinkedIn's Talent Search ranking system, focusing on potential ranking bias across two attributes: *gender* and *race*. To do so, we first construct a dataset of rankings produced by the system, collecting extensive Talent Search results across a diverse set of occupational queries. We then develop a robust labeling pipeline that infers the two demographic attributes of interest for the returned users. To evaluate potential biases in the collected dataset of real-world rankings, we utilize two exposure disparity metrics: deviation from group proportions and MinSkew@k. Our analysis reveals an underrepresentation of minority groups in early ranks across many queries. We further examine potential causes of this disparity, and discuss why they may be difficult or, in some cases, impossible to fully eliminate among the early ranks of queries.

Beyond static metrics, we also investigate the concept of subgroup fairness over time, highlighting *temporal disparities* in exposure and retention, which are often more difficult to audit for in practice. In employer recruiting platforms such as LinkedIn Talent Search, the persistence of a particular candidate over multiple days in the ranking can directly impact the probability that the given candidate is selected for opportunities. Our analysis reveals demographic disparities in this temporal stability, with some groups experiencing greater volatility in their ranked positions than others. We contextualize all our findings alongside LinkedIn's published self-audits of its Talent Search system and reflect on the methodological constraints of a black-box external evaluation, including limited observability and noisy demographic inference. Our work contributes empirical insights and practical guidance for conducting third-party audits of modern socio-technical systems which go beyond the well-studied and standard algorithmic fairness guarantees of predictors.

Code — https://github.com/tina-behzad/LinkedIn-Audit
Extended version — https://arxiv.org/pdf/2511.10752

1 Introduction

LinkedIn is one of the most important platforms for hiring around the world. According to LinkedIn's official statistics, more than 10,000 members worldwide apply for jobs on the platform every minute (LinkedIn 2025). LinkedIn's employer-focused recruiting suite, LinkedIn Recruiter, is one of the platform's most impactful aspects, serving as an influential and widely-used tool through which employers find potential job candidates. LinkedIn's Recruiter platform provides a host of features that connect employer recruiters to relevant candidates for particular roles, via a robust candidate search and screening filters. According to LinkedIn, more than 5.7 million talent professionals across 1.1 million companies use LinkedIn Recruiter to source and hire candidates, and seven people are hired through the platform every minute (LinkedIn Talent Solutions 2025).

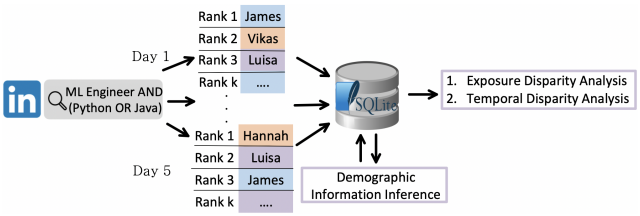

Figure 1: A schematic overview of our pipeline: we issue identical queries to LinkedIn Talent Search over five consecutive days (Section 3.1), ingest the results into our database, and then enrich these records with demographic inferences using external APIs and datasets (Section 3.2). Finally, we carry out exposure-disparity analysis (Section 4) and temporal-disparity analysis (Section 5).

LinkedIn operates at a massive scale, and has a tangible impact on the modern labor market. Given this, the *fairness* of its recruiting platform remains as important an issue as ever. Previous work has demonstrated that some components of social media platforms, such as Meta's employment (Ali et al. 2019; Imana, Korolova, and Heidemann 2021) and education (Imana, Korolova, and Heidemann 2024) ad delivery mechanisms, can be biased. However, to the best of our knowledge, the key aspect of LinkedIn's recruiting platform responsible for directly connecting job market candidates to recruiters, its *recruiter-facing candidate search tool*, has not yet been independently studied.

LinkedIn Talent Search (LTS) allows recruiters to view a *ranking* over individuals when searching for candidates using a particular *query*, which consists of a combination of

*These authors contributed equally.

skills and other criteria specified by the recruiter. For example, a recruiter for a technology company in New York City may construct an LTS query that searches for nearby candidates with skills in Python and at least three years of experience in management.

Given that recruiters have limited resources and place considerable trust in LinkedIn's algorithms,[1] the resulting ranking over candidates for any particular LTS query thus has considerable power in shaping who gets seen, contacted, and ultimately hired. Given LinkedIn's dominance, LTS can influence both individual careers and the diversity of the broader corporate workforce. Minute unfairness or even unintended behaviors on the order of a single-digit percentage can have lasting downstream impacts on thousands of people.

Although LinkedIn's self-reports claim improvements across defined fairness metrics (Geyik, Ambler, and Kenthapadi 2019), independent and replicable external audits are crucial for providing validation of these efforts and improving trustworthiness (Longpre et al. 2025). Our goal is to conduct an independent, replicable external audit of LinkedIn Recruiter's ranking algorithms for potential biases across demographic groups. In addition, we will advance the methodological toolkit (Metaxa et al. 2021) needed for overcoming practical barriers to carry out such audits effectively (Cen and Alur 2024; Casper et al. 2024; Imana, Korolova, and Heidemann 2023).

1.1 Overview of Audit

We conduct an external, fully independent empirical audit of LinkedIn's Recruiter platform candidate search system, LTS, and compare our findings against publicly available LinkedIn's self-reports of its fairness initiatives. First, in Section 2, we elaborate on the challenges of collecting data in the absence of internal or full platform access to LinkedIn Recruiter Search. Next, in Section 3.1, we address the task of data collection and labeling, focusing in particular on the difficulty of inferring demographic information from candidate profiles. In Sections 4 and 5, we discuss and select fairness metrics that capture meaningful notions of equity across multiple time scales and for multiple subgroups, often in the face of incomplete or noisy data. At a high level, our results point to disparities between candidates with differing genders in (1) the early ranks of queries (the first 100 ranks / the first 4 pages of candidate results); and (2) when repeating queries over multiple days. Our work serves both as an analysis of fairness outcomes in LTS and as a reflection on the broader challenges of performing such external audits.

2 LinkedIn's Ranking Pipeline

LinkedIn Recruiter (LTS) operates using a two-stage ranking architecture, a standard approach in large-scale information retrieval systems (Liu et al. 2009; Dang, Bendersky, and Croft 2013). [2] There is a large pool of candidates C. For any particular query q, the platform's goal is to present the recruiter with a ranked list of relevant candidates. Given the scale of LinkedIn's user base, computing a full ranking over all possible candidates in C is prohibitively expensive.

In the first stage, known as candidate retrieval, a small subset of relevant candidates $C_q \subseteq C$ is selected from the overall pool C. The set C_q represents candidates that may be of interest to a recruiter issuing query q. This retrieval stage is powered by *Galene* (Sriram Sankar 2015), LinkedIn's in-house search engine. Galene generates an initial list of candidates based on a feature-level matching utilizing candidate profile aspects such as job titles, skills, employment history, etc.

In the second stage, LinkedIn uses machine-learned models to assign a relevance score to each candidate in C_q. A common paradigm within the second stage is for a score function $r^*(x)$ to map each individual $x \in C_q$ to a *relevance score* in $[0, 1]$; the individuals are then returned to the recruiter in order of decreasing score. One important property of LinkedIn's relevance score is that it captures a combination of two aspects: (1) LinkedIn's estimate that the individual x is qualified for q; *and* (2) LinkedIn's estimate that the candidate is willing to respond to the query q. LinkedIn's motivation for incorporating both components into the relevance score is to satisfy the recruiters: neither unqualified nor uninterested/unmovable candidates would be useful to recruiters, and a large percentage of such candidates would lead to recruiters' dissatisfaction with LTS.[3]

The ranked lists should not only be useful, but also adhere to hiring laws and best practices.[4] Operationalizing hiring laws and best practices is often ambiguous given the uncertainty about how interviewing and hiring laws apply to digital hiring pipelines.[5] Nonetheless, LinkedIn has released information (Geyik, Ambler, and Kenthapadi 2019) about the fairness metric(s) they aim to optimize and some of the interventions that they apply before presenting the final ranked list to the recruiter. We describe one fairness metric and the corresponding post-processing method (DETGREEDY) which, as of 2019, LinkedIn applies to all returned candidate rankings within recruiter search.

Suppose that the overall candidate pool C can be partitioned into m *disjoint* groups $C = g_1 \cup g_2 \cup \cdots \cup g_m$ based on sensitive attributes or profile data. Once the relevant candidate list C_q is obtained, the proportion of each group g_i in query q is defined as $p_i^* = |C_q \cap g_i|/|C_q|$. Notice that

[1] Independent surveys report LinkedIn as the leading social recruiting channel, with over 71% of 1,200 respondents using it (NextThing RPO 2025).

[2] Our description of LinkedIn's talent search is based on the latest publicly available academic and technical publications. We acknowledge that some of these sources are several years old and that the actual systems or algorithms used by LinkedIn may have evolved since their publication.

[3] According to LinkedIn, the "InMail" acceptance rate of candidates accepting such connections from interested recruiters is a key business metric for the company (Ramanath et al. 2018).

[4] We mainly focus on US laws and hiring practices in this work, since the queries we make are conducted from and geographically centered within the US.

[5] See, e.g., the ongoing lawsuit Derek Mobley v. Workday Inc. (Case 3:23-cv-00770).

p_i^* represents the proportion of the *returned relevant candidates* which belong to group g_i, not the proportion of candidates from the entire pool C. Let c_i^k denote the cumulative count of the number of individuals belonging to group g_i from ranks 1 through k (inclusive) in the ranking shown to the recruiter. Geyik, Ambler, and Kenthapadi (2019) propose a suite of *post-processing* algorithms which, for every group g_i, at each $k \in \{1, \ldots, |C_q|\}$, enforce:

$$\lfloor p_i^* \cdot k \rfloor \leq c_i^k \leq \lceil p_i^* \cdot k \rceil.$$

After defining and experimenting with several proposed algorithms in Geyik, Ambler, and Kenthapadi (2019), LinkedIn chose to deploy DETGREEDY in LTS. This algorithm has theoretical guarantees of satisfiability when the number m of groups is at most 3. Since the post-processing algorithm is applied *after* the individuals are sorted by score, individuals are therefore never compared between different groups $g_i \neq g_j$; the scores are only used to rank the individuals *within* their own group. For each position in the ranking, the candidate with the highest score from the currently most underrepresented group is selected. The fact that such an approach works relies crucially on the disjointness of groups; it would be a much more difficult task if the groups were overlapping, as is desired in other fairness notions for candidate ranking (Dwork et al. 2019; Devic et al. 2024).

3 Data Collection

3.1 Retrieving Ranking Data

In this section, we describe how we collected data from LinkedIn's *Recruiter Lite* platform.[6] We note that Recruiter Lite has more limited functionality and access than the full Recruiter platform. However, obtaining access to the full Recruiter platform requires: (1) being a well-established company; (2) a meeting with a LinkedIn representative; and (3) paying an undisclosed amount on the order of many thousands of dollars per recruiter per year — requirements that are virtually impossible to satisfy for independent auditors. In contrast, the Recruiter Lite platform can be accessed with a free trial and can be extended beyond that for $170 a month. Recruiter Lite, to the best of our knowledge and experience, can be easily requested by any real individual with an active and populated LinkedIn profile and a plausible story. As we performed the audit without LinkedIn's express permission or knowledge, and did not have access to any company's full Recruiter platform license, we performed our investigations with Recruiter Lite. Recruiter Lite memberships are subject to several important limitations:

- **Limited search scope**: Search results are restricted to the account owner's 3rd-degree connections. Assuming that each LinkedIn user has around 200 connections,[7] with an estimated 20% overlap in 2nd-degree connections and 30% overlap in 3rd-degree connections, this results in a reachable network of approximately 4.5 million people.
- **Daily candidate limit**: The number of unique candidates viewable in LTS queries is capped at 2,000 per day. This means that collecting a substantial dataset requires multiple days of queries.
- **Result page limit**: Only the first 1,000 candidates (40 pages of 25 candidates each) are viewable for any query.
- **Hidden candidate data**: Candidates outside the account holder's 3rd-degree connections appear in search results as *LinkedIn Member* with anonymized details. Since these candidates are "missing" data, we skip them when scraping. Even if their inclusion were to offset disparities, recruiters using Recruiter Lite would not be able to view or contact these anonymized candidates, so they would not contribute to balancing representation in practice.
- **No access to job-seeking status**: Recruiter Lite membership does not provide visibility into whether a candidate is actively looking for a job by marking themselves as "open to work". This limitation prevents us from tracking transitions between active and passive job-seeking states. These signals could serve as proxies for successful hires or help explain why certain candidates disappear from search results over time.

We employ the Selenium Python package and its built-in Chrome WebDriver to automate data collections (see the full version for additional details). We note that all data collected are publicly available on the LinkedIn Recruiter platform. We collected rankings of around 26,000 candidates across 78 different queries. All queries were collected from the state of New York without a VPN.[8] We also perform many of the same queries across multiple days to investigate the *temporal* aspect of the rankings.

We categorize queries into three levels of search specificity: general queries, general queries with full candidate card information, and position-specific queries with full candidate card information.

Query Set Q_1: General Queries. These queries were selected from the job titles listed by the U.S. Bureau of Labor Statistics (BLS),[9] ensuring that group-level statistics were available for each. In total, we selected 40 titles, aiming to include at least one representative from each major category or subcategory defined by the BLS to preserve diversity across professions. Each query was repeated over five consecutive days. Using these titles as keyword queries, we applied the "NYC Metropolitan Area" location filter and scraped ranked candidates up to rank 200 (eight pages). For each query, we collected all data visible on the candidate cards, including profile name, headline, connection level, profile picture link if available, and background details (education, experience, and skills) without clicking on the "see more" option. We were able to scrape data for up to only eight queries per day. These queries were conducted from December 30th, 2024 to January 28th, 2025. In the following sections we refer to this query set as Q_1.

[6] https://business.linkedin.com/talent-solutions/recruiter-lite

[7] LinkedIn claims that the average number of connections per person in the US was 109 in 2016 (Barbarasa, Barrett, and Goldin 2017).

[8] The Recruiter product could potentially behave differently in different countries; our audit is focused on the US version.

[9] https://www.bls.gov/cps/cpsaat11.htm

Query Set Q_2: General Queries with Full Candidate Card Information. From February 12th to March 28th, 2025, we executed a set of 40 queries. For each query, we collected results up to the 200th rank on five consecutive days within this time period. During this phase, we clicked each candidate card's "See more" button to extract the full list of their experience, background and skills. We refer to this query set as Q_2. While Q_2 contains the same set of queries as Q_1, the key difference is that Q_2 includes full candidate card information. In the remainder of this paper, we focus our analysis on Q_2 for general queries.

Query Sets Q_3 and Q_3': Position Specific Queries with Full Candidate Card Information. To mirror real-world Recruiter searches, we collected descriptions of random LinkedIn job postings (fetched within an Incognito window without logging in to LinkedIn) and prompted OpenAI's o4-mini-high model to generate a query from the posting's requirements. Each query consists of four to five AND-connected terms, where each term itself OR-combines several related skills or job titles. The resulting query returns a focused set of qualified candidates for that job posting. We then iteratively refined each query manually until it returned fewer than $1,000$ qualified candidates, allowing us to retrieve the complete ranked list within the constraints of Recruiter Lite. For every candidate in the ranked lists, we collected the same information as in Set 2, Q_2. For 9 of these queries, data were retrieved on a single day; for the remaining 15, they were collected over five consecutive days. These queries were conducted from March 30th to April 2nd, 2025. In the following sections, we refer to this query set as Q_3. To minimize the impact of missing candidates in the rankings, we selected 8 queries for which the proportion of LinkedinIn members was less than 1% of the total results. Identifying queries that met this criterion was particularly challenging, as LinkedIn frequently returns rankings that include inaccessible members. For each of these queries, we collected the same data for all returned candidates between May 13th and May 23rd, 2025. We refer to this query set as Q_3'.

3.2 Data Labeling

Accurately assessing fairness in algorithmic systems typically requires member-level demographic signals, for example, gender, race, or ethnicity, in order to measure disparities across different groups. For external audits like ours, this means devising a method to infer demographic attributes without direct access to ground truth labels. LinkedIn faces a similar challenge internally as only about 6% of LinkedIn's U.S. members have self-reported their race/ethnicity via LinkedIn's Self-ID survey (Badrinarayanan et al. 2024). To expand fairness assessments beyond this limited group, LinkedIn has developed a Privacy-Preserving Probabilistic Race/Ethnicity Estimation (PPRE) method (Badrinarayanan et al. 2024). This system combines Bayesian Improved Surname Geocoding (BISG, see Elliott et al. (2009)), data from the Self-ID survey, and privacy-enhancing technologies, including secure two-party computation and differential privacy. Crucially, this approach avoids assigning deterministic race/ethnicity labels to individuals and generates probabilistic estimates on the fly, which are encrypted, aggregated, and immediately discarded after measurement. This design ensures that fairness evaluations can be conducted responsibly, without compromising individual privacy or enabling downstream use of inferred demographic attributes.

In the following discussion, we outline the strategies we used to assign gender for candidates. We provide details of our race inference in the full version.

Gender Estimation. In our analysis, we focus on binary gender (male/female).[10] We use candidates' first name to infer their gender using the Social Security Administration's national name dataset (Social Security Administration 2025). For each name, we assign the gender with the higher frequency based on the *Count* column in the dataset, which indicates how many times the name was recorded for each gender. When a name appears with only one associated gender, the assignment is straightforward. For names not found in the dataset, we leverage the free tiers of publicly available commercial APIs (GenderAPI (2025) and GenderizeIO (2025)) to infer gender using the first name. Using the combination of these three sources, we labeled the vast majority of candidates, and only 600 individuals (2.3%) remained unlabeled. To validate the accuracy of our gender labels, two of the authors manually annotated 5,700 candidate profiles using full names and profile pictures (when available).[11] Across the full set of 5,700 manually labeled cases, the agreement with the automatically assigned labels was 95%.

4 Analysis

In this section, we evaluate the effectiveness of potential post-processing methods in LTS using two key metrics that measure disparities in group representation. We focus on gender analysis, with results for racial groups included in the full version. Since both metrics require accurate estimates of each group's overall proportion, we limit our analysis to position-specific queries where the full candidate pool is available (query set Q_3 and Q_3'), allowing for more reliable estimation. We provide the same analysis for the general query set Q_2 in the full version.

4.1 Deviation from Group Proportion

In Geyik, Ambler, and Kenthapadi (2019), LinkedIn noted that they apply the DETGREEDY algorithm (as described in Section 2) using binary gender as the grouping variable. Under this approach, the cumulative count for each group i at rank k should approximately match $p_i^* \cdot k$, where p_i^* represents the share of group i in the overall retrieved candidate pool. In other words, the proportion of candidates from each group should closely reflect p_i^* throughout the ranked list.

For the position-specific queries, Q_3 and Q_3', we can compute observed group proportions directly from the data, and

[10] We acknowledge that this framing does not capture the full spectrum of gender identities, and we recognize the limitations this imposes on the inclusivity and comprehensiveness of our results.

[11] These manual labels were not used in our analysis and were only used to measure the accuracy of our inference.

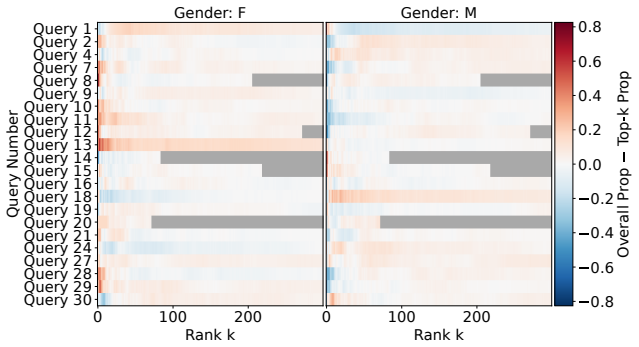

Figure 2: Deviation between the observed top-k gender proportions and the overall candidate pool proportions for the set of queries for which we scraped the full list of returned candidates (Q_3). Each row corresponds to a query, with gender-wise deviations shown across rank positions up to $k = 300$. Gray areas indicate ranks beyond the total number of returned candidates for that query (i.e., the candidate pool was smaller than 300). Red values indicate under-representation relative to the overall group proportion, while blue values indicate over-representation.

use them as an estimate of p_i^* for evaluating exposure disparities by calculating deviation from the true proportion.

The total number of retrieved candidates for queries in Q_3 ranges from 73 to 1,000, the upper limit under our Recruiter Lite membership. We restrict the analysis to queries with less than 15% missing candidates.[12] Here, we examine ranks up to 300 for queries in Q_3. Throughout this analysis and the remainder of Section 4, we rely on the candidate rankings retrieved on the first day each query was run. While these queries provide access to what appears to be the full pool of returned candidates, the presence of even a small number of missing profiles can introduce noise into the estimation of the true group proportions p_i^*. To improve accuracy, we looked for queries with less than 1% missing data and scraped the complete candidate lists for each. We retrieved data for eight such queries (Q_3'), with candidate pool sizes ranging from 111 to 698. Here, and in the subsequent plots, we examine ranks up to 200 for queries in Q_3' since six of the eight queries contain fewer than 200 candidates. The full query text and corresponding candidate pool sizes for both sets are listed in the full version.

Figure 2 and 3 visualize $\left(p_i^* - \frac{n_{i,k}}{k}\right)$ for each query in Q_3 and Q_3', respectively, where $n_{i,k}$ denotes the cumulative count of candidates from group i up to rank k and p_i^* denotes the overall proportion of group i in the candidate pool.

Both plots consistently show that the magnitude of deviation is greatest in the very top ranks and then rapidly diminishes toward zero further down the list.[13] At very low k, the deviation bars (deep red or blue) are largest, but by $k \approx 100$ they have largely collapsed toward white. In other words, small slices of the top of the ranking are where gender skew is most pronounced; as more candidates are being considered, the sample proportions gravitate back toward the overall baseline. For the female panel (left), nearly every query shows a red bias at $k \leq 25$, indicating under-representation of women in the very top slots relative to their overall share. By contrast, the male panel (right) is mostly dominated by blue at the very top.

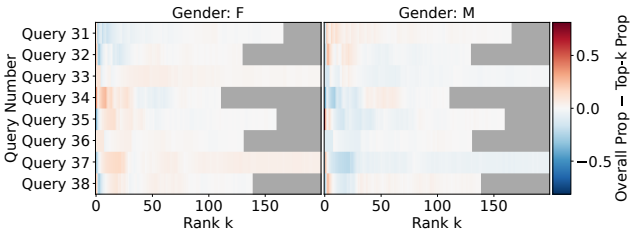

Figure 3: Deviation between the observed top-k gender proportions and the overall candidate pool proportions for the set Q_3' of queries with less than 1% missing members. Gender-wise deviations are shown across rank positions up to $k = 200$.

After about $k \geq 75$, almost all entries are near zero, so by the third or fourth page of results, there is essentially no gender skew. However, this does not diminish the impact of disparities observed at the top of the ranking. Algorithmic systems are subject to position and trust biases; users are more likely to engage with candidates near the top and to trust that those candidates are the most qualified (Joachims et al. 2007, 2005). In platforms like LinkedIn Recruiter, where recruiters typically begin reviewing from the top of the list and may never reach lower-ranked pages, skew at early ranks carries the most weight. Even small disparities at the top can translate into meaningful inequities in visibility and opportunity.

4.2 Skew@k

Although computing deviation from the expected group proportions at top-k ranks is a standard approach to measure statistical parity, and aligns with standard metrics such as Normalized Discounted Difference proposed by Yang and Stoyanovich (2017), in the work Geyik, Ambler, and Kenthapadi (2019), LinkedIn introduced an alternative metric known as *MinSkew@k*.

This metric quantifies the worst-case deviation from the target representation among all protected groups at a given cutoff k. Specifically, the skew for group g_i in the ranked list τ_r is defined as:

$$\text{Skew}_{g_i}@k(\tau_r) = \log\left(\frac{p_{k,g_i}^{\tau_r}}{p_{g_i}^{q}}\right), \quad (1)$$

where $p_{k,g_i}^{\tau_r}$ denotes the proportion of candidates with attribute value g_i in the top-k of the ranked list τ_r, and $p_{g_i}^{q}$ is

[12] The missing candidate rate thresholds for each set were chosen to balance data quality with maintaining adequate query coverage. Recall that candidates are "missing" in the Recruiter Lite rankings if they are further than three hops in the connection graph from the recruiter (Section 3.1).

[13] We use the terms top ranks and early ranks interchangeably throughout the paper, both referring to candidates appearing at the beginning of the rankings.

the desired (target) proportion for g_i in the given query q. Subsequently, MinSkew@k is defined as:

$$\text{MinSkew@}k(\tau_r) = \min_{g_i \in G} \text{Skew}_{g_i}@k(\tau_r), \qquad (2)$$

capturing the most disadvantaged group's deviation at rank k. According to the reported results of Geyik, Ambler, and Kenthapadi (2019), the average MinSkew@100 improves from -0.259 to -0.011 after applying the DETGREEDY algorithm. They also report similar improvements across other cutoffs, with MinSkew@k values approaching zero consistently for over 95% of the queries (see the full version for a step-by-step example of how this metric is calculated and how it varies as the underlying distribution changes.).

The logarithmic nature of the metric makes it highly sensitive to small discrepancies in observed vs. expected proportions. While this makes it powerful for detecting even mild imbalances, it also means that inaccurate or noisy estimates of p^*, especially in external audits like ours, can lead to misleading skew values. Accurate estimation of the target distribution is therefore essential for the meaningful application of this metric. Therefore, for the following comparison, we focus on the subset of queries with less than 1% missing candidates from the set where the full candidate pool was retrieved ($Q_3 \cup Q'_3$). This results in 12 queries. Figure 4 displays the Skew@k metric for both gender groups across the 12 queries, evaluated up to rank 200. The top panel shows the skew for women, while the bottom panel shows the skew for men. Note that the y-axis scales differ between the two plots: female skew values range from approximately -1.5 to $+1.5$, while male skew values are mostly bounded between -0.4 and $+0.4$. This indicates that deviations from expected representation are more extreme and variable for women.

We also checked for any correlation between Skew@k and group size (overall proportion) and observed no patterns (see the full version). Across queries, we observe that Skew@k for women tends to be sharply negative at early ranks, indicating significant under-representation at the top of the list. For many queries, skew gradually approaches zero as k increases, consistent with DETGREEDY's goal of aligning cumulative group representation with expected proportions. However, noticeable dips and peaks occur at rank intervals that align with page boundaries in LinkedIn Recruiter (e.g., at $k = 25, 50, 75$), which could be due to certain applied ranking-metric optimizations taking place at these cut-offs.

A potential confounder in interpreting the Skew@k metric at the very top of the ranking is the discrete nature of candidate placements. For small values of k, only a few proportions are actually attainable. For example, when $k = 3$, one can only realize 33%, 66%, or 100% representation. Even in an optimally balanced ranking it might be impossible to fully eliminate skew among the early ranks of queries. To correct for this, we compute, at each rank cutoff k, the "unavoidable" skew, and then subtract that baseline from our observed skew. The resulting deviation isolates true over- or under-representation beyond what is imposed by integral candidate placement. Our results show that even after subtracting the best attainable skew at each cutoff, the female candidates' skew at the very top ranks remains well below the reported 0.011 threshold (see the full version for the baseline-corrected skew analysis). Furthermore, for every cutoff k, the female group consistently exhibits the lowest possible skew across these queries and therefore derives the MinSkew@k value, a consistency that itself could indicate a potential underlying discrepancy in representation.

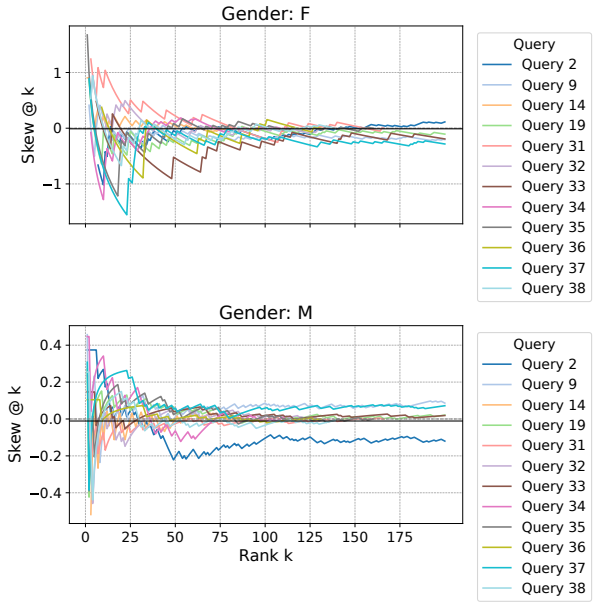

Figure 4: Skew@k for each gender across 12 queries with less than 1% missing candidates, evaluated up to rank 200. The top plot shows skew values for women; the bottom plot shows skew values for men. The black horizontal line indicates the average MinSkew@100 value of -0.011 reported in Geyik, Ambler, and Kenthapadi (2019).

To assess whether the average MinSkew deviates from the reported benchmark of -0.011 after accounting for both between-query variability and day-to-day noise, we fitted intercept-only linear mixed-effects models at page cutoffs $k \in \{25, 50, 75, 100\}$, with a random intercept for each query (see the full version for details). In every case through page 4 ($k = 100$), the Wald tests reject $H_0 : E[\text{MinSkew@}k] = -0.011$ with $p < 0.001$. This confirms that the observed MinSkew is significantly more negative than -0.011 and cannot be explained by day-to-day or query-to-query noise alone. This corroborates our findings in Section 4.1 which point to group disparities in the early (before 100) ranks. Both metrics consistently highlight that disparities are most severe in the top portion of the list and mostly in favor of the male group.

5 Temporal Aspects of Fairness in Ranked Candidate Lists

LinkedIn Recruiter operates as a two-sided marketplace, where recruiters and job seekers have distinct goals and preferences. On one side, recruiters performing repeated searches for the same role expect fresh results each day, ideally surfacing new candidates they have not yet reviewed. On the other side, candidates who are relevant to a given search

would reasonably expect to appear consistently in the results across time, ensuring continued visibility.

The LTS system must balance these competing objectives while treating all demographic groups equitably. In particular, there should be no systematic differences in candidate retention across days based on gender, race, or other protected attributes. Temporal discrepancies in exposure can create unequal visibility and opportunity, even when single-day rankings appear fair in isolation.

Temporal fairness, and more specifically subgroup retention across repeated queries, is an underexplored dimension in the literature (Patro et al. 2022; Liu et al. 2018), and one that, despite its importance, is not explicitly mentioned in any of the LinkedIn public-facing communications. To evaluate subgroup stability over time, we define the *churn rate* for group g_i between two rankings (r_t) of the same query on start day s and end day e, over the top k results as:

$$\text{Churn}_{g_i}^{s\to e}(k) = \frac{|\{x \in g_i \,|\, x \in \text{Top-}k(r_s) \wedge x \notin \text{Top-}k(r_e)\}|}{|\{x \in g_i \,|\, x \in \text{Top-}k(r_s)\}|}$$

This metric captures the proportion of candidates in group g_i who appeared in the top k results at time s but were no longer present at time e. While the overall level of churn, whether high or low, may reflect deliberate design choices (e.g., promoting freshness vs. stability), it is important for the churn rate to be approximately consistent across demographic groups. This is due to the fact that large disparities in churn can lead to unequal exposure over time, which in turn can create disparate candidate outcomes.

The left and right panels of Figure 5, for females and males respectively, show $\text{Churn}_{g_i}^{1\to j}(k)$ for $k \in \{25, 50, \ldots, 200\}$ and $j \in \{2, 3, 4, 5\}$ in the set $Q_3 \cup Q_3'$ of position specific queries with less than 15% missing candidates and 5 consecutive days of data. Higher churn rate, indicating a greater proportion of individuals dropping out of the top-k ranking, is represented using a darker shade. Recall, though, that our pipeline does not allow us to determine whether departures are due to hires or reshuffling.

Overall, we observe that churn rates are highest at the very top and then steadily decline as k increases, reflecting the competition for top slots and a relative stabilization deeper in the list. This pattern holds true for both genders.

At $k = 25$ and $k = 50$, women churn about 0.07 units more than men on average across days, indicating a less stable presence in the top-k pools. Male drop-outs follow a more predictable pattern, with smaller turnover from day 1→2 than from day 1→3, 4, or 5, while women's exits are more erratic, suggesting greater volatility. We examined whether churn correlates with overall group representation and found no association. Using mixed-effects models, Wald tests reveal statistically significant group differences in churn at $k = 25$ and $k = 50$. Additional query groups, detailed results, and methodological details appear in the full version.

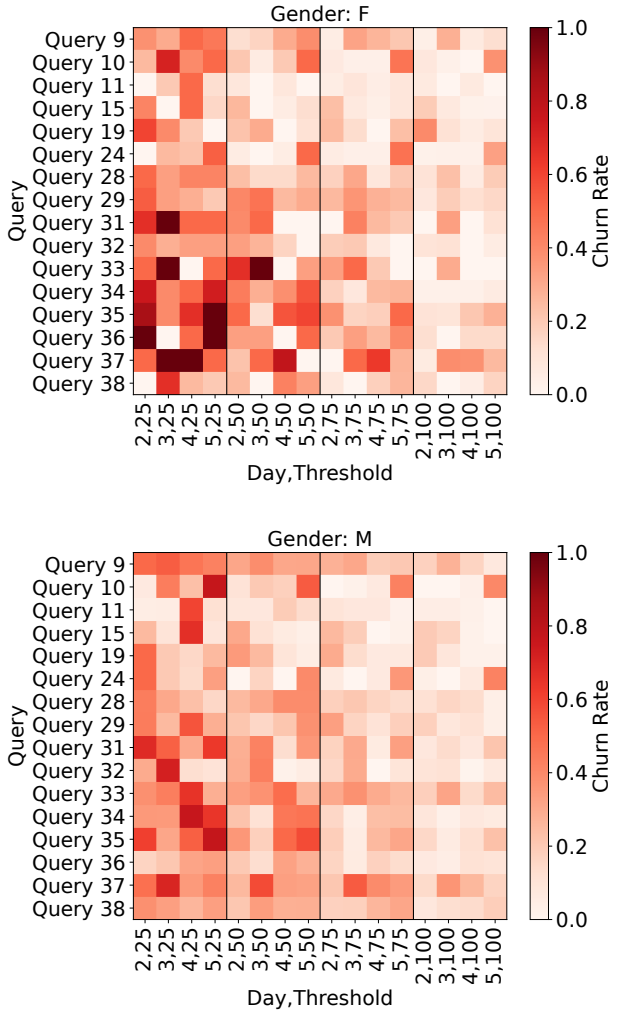

Figure 5: (**Left**): Heat map of day-to-day churn rates $\text{Churn}_{g_i}^{1\to j}(k)$ for the female candidates across position-specific queries (with less than 15% missing candidates) over five consecutive days, evaluated at top-k cutoffs $k \in \{25, 50, \ldots, 100\}$. Darker shades indicate higher proportions of candidates dropping out of the top k. (**Right**): Identical plot for male candidates.

6 Conclusion

We conducted an independent external audit of LinkedIn Recruiter's ranking algorithms, examining disparities in candidate representation across gender and racial groups. Our results suggest the use of demographic-aware post-processing, as disparities decrease at lower ranks but persist near the top. We also find temporal instability, with churn rates varying across groups. Beyond these findings, we provide methodological guidance for independent audits of platforms with restricted access; limitations are discussed in the full version.

Acknowledgments

Our study received IRB exemption UP-24-01124 at the University of Southern California. This work was supported by multiple National Science Foundation grants: CNS-1956435, CNS-2344925, and NSF CAREER Award CCF-2239265. V. Sharan was also supported by an Okawa Foundation Award, and A. Korolova by the Alfred P. Sloan Research Fellowship.

We thank the anonymous AAAI reviewers for helpful comments and feedback used to improve the work.

References

Ali, M.; Sapiezynski, P.; Bogen, M.; Korolova, A.; Mislove, A.; and Rieke, A. 2019. Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, volume 3, 1–30. ACM.

Badrinarayanan, S.; Osoba, O.; Cheng, M.; Rogers, R.; Jain, S.; Tandra, R.; and Pillai, N. S. 2024. Privacy-Preserving Race/Ethnicity Estimation for Algorithmic Bias Measurement in the US. *arXiv preprint arXiv:2409.04652*.

Barbarasa, E.; Barrett, J.; and Goldin, N. 2017. Skills gap or signaling gap?: Insights from LinkedIn in emerging markets of Brazil, India, Indonesia, and South Africa. *Solutions for Youth Employment*.

Casper, S.; Ezell, C.; Siegmann, C.; Kolt, N.; Curtis, T. L.; Bucknall, B.; Haupt, A.; Wei, K.; Scheurer, J.; Hobbhahn, M.; et al. 2024. Black-box access is insufficient for rigorous ai audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2254–2272.

Cen, S. H.; and Alur, R. 2024. From Transparency to Accountability and Back: A Discussion of Access and Evidence in AI Auditing. In *EAAMO '24*. New York, NY, USA: Association for Computing Machinery. ISBN 9798400712227.

Dang, V.; Bendersky, M.; and Croft, W. B. 2013. Two-stage learning to rank for information retrieval. In *European Conference on Information Retrieval*, 423–434. Springer.

Devic, S.; Korolova, A.; Kempe, D.; and Sharan, V. 2024. Stability and Multigroup Fairness in Ranking with Uncertain Predictions. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Dwork, C.; Kim, M. P.; Reingold, O.; Rothblum, G. N.; and Yona, G. 2019. Learning from outcomes: Evidence-based rankings. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, 106–125. IEEE.

Elliott, M. N.; Morrison, P. A.; Fremont, A.; McCaffrey, D. F.; Pantoja, P.; and Lurie, N. 2009. Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, 9(2): 69–83.

GenderAPI. 2025. GenderAPI: Gender Detection API. https://www.genderapi.io/. Accessed: 2025-04-23.

GenderizeIO. 2025. Genderize.io: Predicting Gender from Names. https://genderize.io/. Accessed: 2025-04-23.

Geyik, S. C.; Ambler, S.; and Kenthapadi, K. 2019. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*, 2221–2231.

Imana, B.; Korolova, A.; and Heidemann, J. 2021. Auditing for Discrimination in Algorithms Delivering Job Ads. In *The Web Conference (WWW)*.

Imana, B.; Korolova, A.; and Heidemann, J. 2023. Having your Privacy Cake and Eating it Too: Platform-supported Auditing of Social Media Algorithms for Public Interest. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–33.

Imana, B.; Korolova, A.; and Heidemann, J. 2024. Auditing for racial discrimination in the delivery of education ads. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2348–2361.

Joachims, T.; Granka, L.; Pan, B.; Hembrooke, H.; and Gay, G. 2005. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, 154–161. New York, NY, USA: Association for Computing Machinery. ISBN 1595930345.

Joachims, T.; Granka, L.; Pan, B.; Hembrooke, H.; Radlinski, F.; and Gay, G. 2007. Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. In *ACM Transactions on Information Systems*, volume 25, 7–es. New York, NY, USA: Association for Computing Machinery.

LinkedIn. 2025. About LinkedIn: Company Statistics. Accessed: 2025-07-25.

LinkedIn Talent Solutions. 2025. Talent Solutions: Tools for Hiring and Recruiting. Accessed: 2025-07-25.

Liu, L. T.; Dean, S.; Rolf, E.; Simchowitz, M.; and Hardt, M. 2018. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, 3150–3158. PMLR.

Liu, T.-Y.; et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3): 225–331.

Longpre, S.; Klyman, K.; Appel, R. E.; Kapoor, S.; Bommasani, R.; Sahar, M.; McGregor, S.; Ghosh, A.; Blili-Hamelin, B.; Butters, N.; et al. 2025. In-house evaluation is not enough: Towards robust third-party flaw disclosure for general-purpose ai. *arXiv preprint arXiv:2503.16861*.

Metaxa, D.; Park, J. S.; Robertson, R. E.; Karahalios, K.; Wilson, C.; Hancock, J.; Sandvig, C.; et al. 2021. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction*, 14(4): 272–344.

NextThing RPO. 2025. Employ Recruiter Nation Report 2024: Empowering People-First Recruiting. Technical report, NextThing RPO. Accessed: 2025-08-01.

Patro, G. K.; Porcaro, L.; Mitchell, L.; Zhang, Q.; Zehlike, M.; and Garg, N. 2022. Fair ranking: a critical review, challenges, and future directions. In *Proceedings of the 2022*

ACM conference on fairness, accountability, and transparency, 1929–1942.

Ramanath, R.; Inan, H.; Polatkan, G.; Hu, B.; Guo, Q.; Ozcaglar, C.; Wu, X.; Kenthapadi, K.; and Geyik, S. C. 2018. Towards Deep and Representation Learning for Talent Search at LinkedIn. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, 2253–2261. New York, NY, USA: Association for Computing Machinery. ISBN 9781450360142.

Social Security Administration. 2025. Baby Names Database. https://www.ssa.gov/oact/babynames/. Accessed: 2025-02-10.

Sriram Sankar, A. M. 2015. Did You Mean Galene? https://engineering.linkedin.com/search/did-you-mean-galene. Accessed: 2025-07-08.

Yang, K.; and Stoyanovich, J. 2017. Measuring Fairness in Ranked Outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, SSDBM '17. New York, NY, USA: Association for Computing Machinery. ISBN 9781450352826.