A Discrimination Report Card

Patrick Kline, UC Berkeley
Evan K. Rose, University of Chicago
Christopher Walters, UC Berkeley

Goh Keng Swee Seminar (NUS)

August 17th, 2023

Who discriminates?

- Increasing agreement that wage setting conduct varies systematically across **firms** (Card et al., 2018). What about *recruiting* conduct?
- Large literature uses correspondence studies to measure market-average discrimination against these protected characteristics (Bertrand and Duflo, 2017)
- Little known about discriminatory conduct of specific employers despite widespread interest from the public
  - Firms: advertise commitment to diversity; spend on Chief Diversity Officers, DEI consultants, HR intermediaries.
  - Job seekers: consult crowd-sourced data (e.g., Glassdoor) on inclusivity.
  - Enforcement agencies: EEOC Systemic Unit focuses on cases with broad impact. OFCCP audits fed contractors for compliance w/ EEO laws.

Measuring employer-level discrimination

- Recent work uses correspondence experiments combined with empirical Bayes and large-scale inference methods to study discrimination by particular employers
- Kline and Walters (2021): Reanalysis of several correspondence experiments
  - Framework: Correspondence study as ensemble of job-specific micro-experiments, each with its own response probabilities
  - Key findings: Tremendous heterogeneity in discrimination across jobs; possible to detect discrimination at some individual jobs with high confidence
- Kline, Rose, and Walters (2022): Massive correspondence experiment of discrimination at 108 large firms
  - 1,000 applications sent to 100+ jobs at each company
  - Signaled race/gender with distinctive names
  - Key finding 1: Wide variation across firms in bias against Black / female names; top 20% account for $\sim$50% of total
  - Key finding 2: Half of variation across firms explained by two-digit industry

Summarizing firm-level conduct

- Experimental results demonstrate that discrimination is highly concentrated in a small set of employers, but estimate for any given employer may be subject to substantial sampling error
- How should we communicate information on firm-specific discrimination to a broad audience?
  - Scientific communication generally aided by transparency (Andrews and Shapiro, 2021)
  - But some audiences may find it difficult to interpret complex statistical evidence (Mullainathan, 2002; Mullainathan et al., 2008; Bordalo et al., 2016)
- Scholars and policymakers increasingly construct simple "report cards" summarizing econometric estimates of quality for various institutions: colleges (Chetty et al., 2017), K-12 schools (Bergman et al., 2020; Angrist et al., 2021), teachers (Bergman and Hill, 2018; Pope, 2019), healthcare providers (Brook et al., 2002; Pope, 2009), neighborhoods (Chetty and Hendren, 2018; Chetty et al., 2018)

Today's agenda: discrimination report cards

- An Empirical Bayes report card that grades the discriminatory conduct of firms
- Report card scheme formalizes tradeoff between informativeness and reliability
  - Audience makes pairwise inferences on relative discrimination based on grades
  - Combine EB posterior pairwise ranking probabilities to construct a global partial ordering
  - Asymmetric preferences over correct rankings vs. mistakes $\mapsto$ optimal coarsening with few grades
  - Analogue of False Discovery Rates for summarizing grade reliability

Related literature

- Audit and correspondence experiments for measuring racial discrimination (Daniel, 1968; Wienk et al., 1979; Heckman and Siegelman, 1993; Heckman, 1998; Bertrand and Mullainathan, 2004; Pager et al., 2009; Nunley et al., 2015; Bertrand and Duflo, 2017; Quillian et al, 2017; Baert, 2018; Gaddis, 2018; Neumark, 2018; Kline, Rose, and Walters, 2022)
- Scientific communication (Savage, 1954; Andrews and Shapiro, 2021; Viviano, Wuthrich, Niehaus, 2021; Korting et al., 2021)
- Limited attention / signal coarsening (Mullainathan, Schwartzstein, and Shleifer, 2008; Pope, 2009; Gilbert et al., 2012; Lacetera, Pope, and Sydnor, 2012; Sejas-Portillo et al., 2020)
- Empirical Bayes inference / selection rules / false discovery rates (Robbins, 1964; Benjamini and Hochberg, 1995; Efron et al., 2001; Storey, 2002; Armstrong, 2015; Efron, 2016; Armstrong, Kolesár, Plagborg-Møller, 2020; Kline and Walters, 2021; Gu and Koenker, 2023)
- Econometrics of ranks (Portnoy, 1982; Berger and Deely, 1988; Laird and Louis, 1989; Sobel, 1993; Mogstad et al., 2020; Andrews et al., 2021; Gu and Koenker, 2022)
- Social choice / vote aggregation (Borda, 1784; Condorcet, 1785; Kemeny, 1959; Smith, 1973; Young and Levenglick, 1978; Young, 1986)

Experimental design

Sampling frame (I/II)

Holding companies split into brands with separate hiring portals (e.g., Berkshire Hathaway into Geico, McLane, Fruit of the Loom, etc.)

Fortune 500

InfoGroup and Burning Glass data merged to measure geographic distribution of establishments and vacancies

123 firms with sufficient expected geographic scope

Hiring platforms investigated to test for feasibility of submitting fictitious applications

108 feasible to audit

Sampling frame (I/II)

Holding companies split into brands with separate hiring portals (e.g., Berkshire Hathaway into Geico, McLane, Fruit of the Loom, etc.)	Fortune 500
InfoGroup and Burning Glass data merged to measure geographic distribution of establishments and vacancies	123 firms with sufficient expected geographic scope
Hiring platforms investigated to test for feasibility of submitting fictitious applications	108 feasible to audit

Compustat: U.S. employment at 108 sampled firms totaled $\sim$**15M** in 2020

Sampling frame (II/II)

4 not sampled in wave 1 due to COVID interruption; 9 firms dropped before completion due to technological constraints; 19 added in wave 2 or later; 4 posted insufficient jobs to sample in all waves

72 sampled in all waves

36 sampled in subset of waves

Job sampled from universe of entry-level vacancies posted on each firm's hiring portal; most recently posted job prioritized

25 vacancies in distinct counties sampled each wave

One pair of applications (1 black and 1 white name) sent every 1-2 days; gender (50% male), age (uniform age 20-60), gender identity (5% gender-neutral, 5% same-gender pronouns), and sexual orientation (10% LGBTQ student club, 10% other club) unconditionally randomly assigned

8 applications sent to each vacancy

Resume characteristics

Job applications manipulate employer perceptions of several protected characteristics:

- Race & gender: distinctive first names obtained from Bertrand and Mullainathan (2004) + NC data on speeding tickets. Last names from Census
- Age: year of high school graduation

Stratify on race (4B/4W), unconditional random assignment of gender, age, as well as LGBTQ affiliation and gender identity

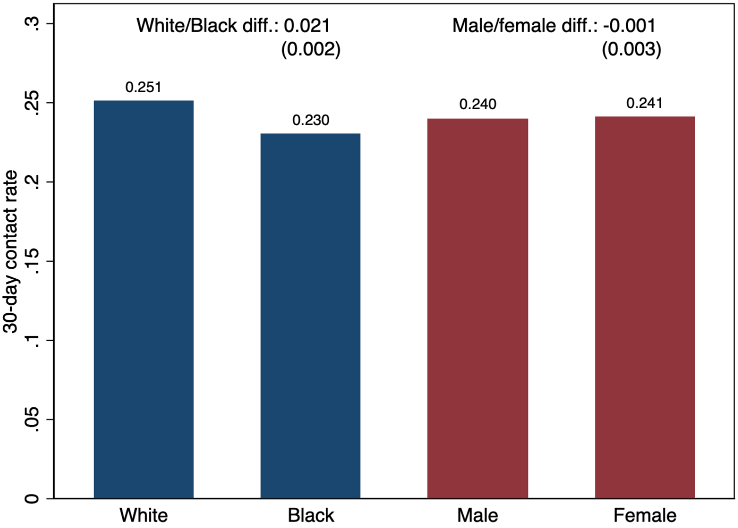
Random assignment of job-appropriate experience, high school, associate degree, resume design, answers to personality tests, etc.

Fully automated sampling of vacancies and submission of apps

Summary stats

	A. All firms			B. Balanced sample		
	White	Black	Difference	White	Black	Difference
Resume characteristics						
Female	0.499	0.499	-0.001	0.500	0.498	0.003
Over 40	0.535	0.535	0.000	0.534	0.533	0.002
LGBTQ club member	0.081	0.082	-0.001	0.079	0.080	-0.001
Academic club	0.040	0.042	-0.002	0.039	0.042	-0.003*
Political club	0.042	0.042	0.001	0.042	0.041	0.001
Gender-neutral pronouns	0.041	0.041	-0.001	0.040	0.040	0.000
Same-gender pronouns	0.043	0.042	0.001	0.042	0.041	0.001
Associate degree	0.476	0.485	-0.009**	0.478	0.485	-0.006*
N applications	41837	41806	83643	32703	32665	65368
N jobs			11114			8667
N firms			108			72
1/2/3/4/5 waves			3/4/15/16/72			

Mean differences: White names favored by 2.1pp, zero average gender difference



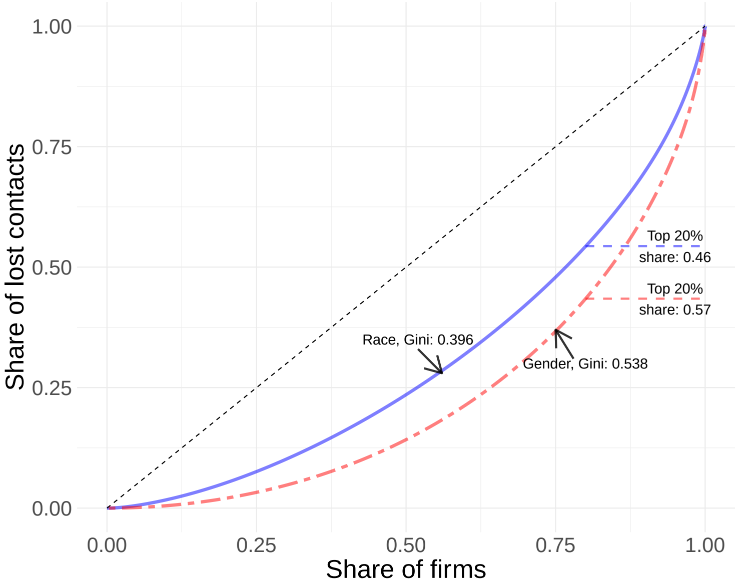
Std. devs.: Substantial heterogeneity across firms for both race and gender

Estimates of firm heterogeneity in race and gender discrimination

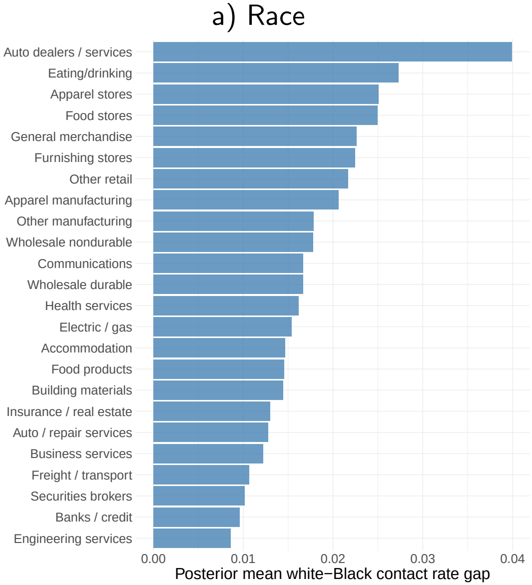
	Mean contact gap (1)	Bias-corrected std. dev. of contact gaps (2)
Race (White - Black)	0.021	0.0185
	(0.002)	(0.0031)
Gender (Male - Female)	-0.001	0.0267
	(0.003)	(0.0038)

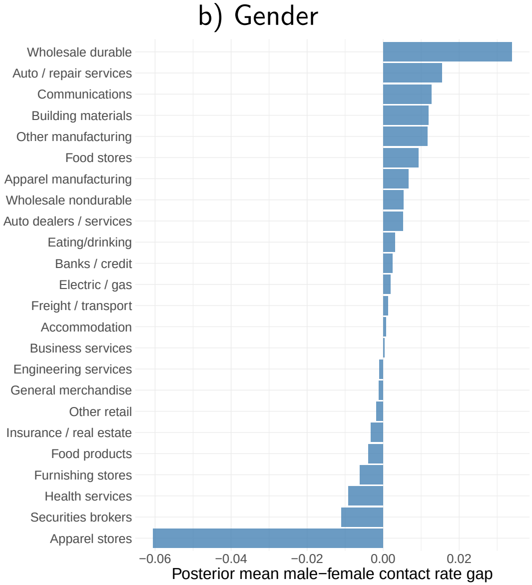
Estimates from Kline, Rose, and Walters (2022).

Lorenz curves: Top 20% of firms explain ∼50-60% of lost contacts



Posterior mean gaps by industry





A Discrimination Report Card

Preliminaries

- n firms, indexed by $i \in \{1, \dots, n\} \equiv [n]$
- Discrimination at firm i parameterized by $\theta_i \in \mathbb{R}$ (proportional contact gap)
- For each firm observe: $Y_i = (\hat{\theta}_i, s_i)$
- $\{Y_i\}_{i=1}^n$ mutually independent conditional on $\theta = (\theta_1, \dots, \theta_n)'$
- Large sample approximation

$$\hat{\theta}_i \mid \theta_i, s_i \sim \mathcal{N}(\theta_i, s_i^2)$$

Gambling over contrasts

Suppose smooth *i.i.d.* prior G over $\{\theta_i\}_{i\in[n]}$ and consider the following risky gamble:

- Observe realizations (y_i, y_j) of (Y_i, Y_j)
- Propose partial ordering $d = (d_i, d_j) \in \{1,2\}^2$ of θ_i and θ_j
- If ordering correct: payoff $= \lambda \in (0,1]$
- If ordering incorrect: payoff = -1
- Declare a tie / abstain: payoff $= 0$

Given posterior $\pi_{ij} = \Pr_G(\theta_i > \theta_j | Y_i = y_i, Y_j = y_j)$, expected utility of choosing d is

$$EU(\pi_{ij}, d) = \underbrace{[\lambda\pi_{ij} - (1-\pi_{ij})]}_{(1+\lambda)\pi_{ij}-1} \cdot 1\{d_i > d_j\} + \underbrace{[\lambda(1-\pi_{ij}) - \pi_{ij}]}_{(1+\lambda)(1-\pi_{ij})-1} \cdot 1\{d_i < d_j\}$$

Optimal decision

Maximize EU with posterior threshold rule:

- Set $d_i > d_j$ iff $\pi_{ij} > \frac{1}{1+\lambda}$
- Set $d_i < d_j$ iff $1 - \pi_{ij} > \frac{1}{1+\lambda}$
- Otherwise set $d_i = d_j$

Threshold approaches 1 as $\lambda \to 0$, yielding all ties

No ties when $\lambda = 1$ bc threshold is $1/2$ (and smooth prior)

Pooling pairs

Now consider all $\binom{n}{2}$ firm pairs. Loss of grades $d = (d_1, \ldots, d_n)' \in [n]^n$ is:

$$L(\theta, d; \lambda) = \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i} \Bigg[\underbrace{1\{\theta_i > \theta_j, d_i < d_j\} + 1\{\theta_i < \theta_j, d_i > d_j\}}_{\text{discordant pairs}} - \lambda \Bigg(\underbrace{1\{\theta_i < \theta_j, d_i < d_j\} + 1\{\theta_i > \theta_j, d_i > d_j\}}_{\text{concordant pairs}} \Bigg) \Bigg]$$

Note: when $\lambda = 1$, loss is the negative of Kendall (1938)'s tau coefficient between d and θ, i.e., bubble-sort distance

Quantifying mistakes

Define the *Discordance Proportion* as

$$DP(\theta, d) = \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i} [1\{\theta_i > \theta_j, d_i < d_j\} + 1\{\theta_i < \theta_j, d_i > d_j\}]$$
$$= \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i} |1\{\theta_i > \theta_j\} - 1\{d_i > d_j\}| \cdot 1\{d_i \neq d_j\}$$

- *DP* measures frequency of misrankings (Type III error rate)
- Can limit by coarsening grades / declaring ties

Too much information

Letting $\tau(\theta, d) \in [-1, 1]$ denote Kendall's tau, we can write the loss

$$L(\theta, d; \lambda) = (1 - \lambda) \, DP(\theta, d) - \lambda \tau(\theta, d)$$

- Parameter λ governs trade-off between information content of rankings (τ) and mistake frequency (DP)
- $1 - \lambda$ measures *discordance aversion*
- When $\lambda < 1$, willing to coarsen grades to avoid discordances

Optimal grades

The Bayes risk of a fixed vector of grades d given data realization y is

$$
\begin{aligned}
\mathcal{R}(\pi, d; \lambda) &= \mathbb{E}_G[L(\theta, d; \lambda) | Y = y] \\
&= \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i} \Bigg[(1 - \pi_{ij})\, 1\{d_i > d_j\} + \pi_{ij} 1\{d_i < d_j\} \\
&\quad - \lambda\, (1 - \pi_{ij})\, 1\{d_i < d_j\} - \lambda \pi_{ij} 1\{d_i > d_j\} \Bigg]
\end{aligned}
$$

Optimal grades are

$$
d^*(\lambda) = \arg\min_{d \in [n]^n} \mathcal{R}(\pi, d; \lambda)
$$

Condorcet paradox

While objective $\mathcal{R}(\pi, d; \lambda)$ is separable across pairs, logical constraints prevent pairwise optimization via comparing π_{ij} to threshold $(1+\lambda)^{-1}$

Example (Three firms, normal posteriors)

Suppose $\theta_i | Y_i = y_i \sim N(\mu_i, 1)$. Then if θs are independent:

$$\pi_{ij} = Pr(\theta_i > \theta_j | Y_i = y_i, Y_j = y_j) = \Phi\left(\frac{\mu_i - \mu_j}{\sqrt{2}}\right)$$

- Let $\lambda = 1/4 \implies (1+\lambda)^{-1} = 0.8$
- Suppose $(\mu_1, \mu_3) = (2, 0)$, so that $\pi_{13} = \Phi(\sqrt{2}) = .92$ and $\pi_{31} = 1 - \pi_{13} = .08$
- Then it is optimal to rank $\theta_1 > \theta_3$.
- But if $\mu_2 \in (0.81, 1.19)$, rank (θ_1, θ_2), (θ_2, θ_3) as ties because $\max\{\pi_{12}, \pi_{23}\} < 0.8$

This is a logical contradiction violating **transitivity**

ILP formulation

Define indicators $d_{ij} = 1\{d_i > d_j\}$ and $e_{ij} = 1\{d_i = d_j\}$. We can rewrite our problem as choosing $\{d_{ij}, e_{ij}\}_{i<j\leq n}$ to minimize

$$\sum_{i=2}^{n}\sum_{j=1}^{i}\left[(1-\pi_{ij})\, d_{ij} + \pi_{ij}\,(1-e_{ij}-d_{ij}) - \lambda\,(1-\pi_{ij})\,(1-e_{ij}-d_{ij}) - \lambda\pi_{ij}d_{ij}\right]$$

s.t. to the following transitivity constraints on any triple $(i,j,k) \in [n]^3$:

$$d_{ij} + d_{jk} \leq 1 + d_{ik}, \quad d_{ik} + (1 - d_{jk}) \leq 1 + d_{ij}, \quad e_{ij} + e_{jk} \leq 1 + e_{ik}$$

Linear objective + linear constraints $\Longrightarrow$ **integer linear programming**

A connection to social choice

When $\lambda = 1$ we seek to minimize

$$\sum_{i=2}^{n} \sum_{j=1}^{i} (2\pi_{ij} - 1)(d_{ji} - d_{ij})$$

If π_{ij} is viewed as the number of votes for $\theta_i > \theta_j$ the constrained minimizer $d^*(1)$ of this objective is the Kemeny - Young voting method (aka Condorcet's rule)

Young (1988) showed that $d^*(1)$ is

- The most likely ranking (aka the maximum likelihood estimator) when all voters have a common probability $> 1/2$ of deciding pairwise contrasts correctly
- The unique ranking rule that is anonymous, neutral, unanimous, and satisfies reinforcement and independence of remote alternatives

Condorcet property

Condorcet criterion: if there is a unit i that wins pairwise election against all $j \neq i$, then i will be top ranked.

Theorem (λ-Condorcet Criterion)

Suppose that firm i satisfies $\pi_{ij} > (1+\lambda)^{-1}\ \forall\ j \neq i$. Then $d_i > d_j\ \forall\ j \neq i$.

Moreover, suppose that firm k satisfies $\pi_{ik} > (1+\lambda)^{-1}$ and $\pi_{kj} > (1+\lambda)^{-1}\ \forall\ j \neq i, j \neq k$, then $d_i > d_k > d_j\ \forall\ j \neq i, j \neq k$.

- Equivalent argument yields selection of bottom ranked "losers."
- With $\lambda < 1$, ties emerge. Show in paper that λ-ranking scheme selects notion corresponding to Smith (1973) set.

Discordance Rates

Define the Discordance Rate (DR) as the expected DP of optimal grades:

$$DR(\lambda) \quad = \quad \binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i} 1\left\{d_i^*(\lambda) < d_j^*(\lambda)\right\} \pi_{ij} + 1\left\{d_i^*(\lambda) > d_j^*(\lambda)\right\}(1-\pi_{ij}).$$

The DR between a specific pair of grades g and $g' < g$ is

$$DR_{g,g'}(\lambda) = \frac{\sum_{i=1}^{n} \sum_{j \neq i} 1\left\{d_i^*(\lambda) = g\right\} 1\left\{d_j^*(\lambda) = g'\right\}(1-\pi_{ij})}{\sum_{i=1}^{n} \sum_{j \neq i} 1\left\{d_i^*(\lambda) = g\right\} 1\left\{d_j^*(\lambda) = g'\right\}}.$$

- $DR_{g,g'}$ analogous to False Discovery Rate of collection of 1-sided contrasts
- DR decomposes into weighted average of the $\{DR_{g,g'}\}$ and $DR_{g,g} = 0$

Empirics: Names

Estimated R^2 of race and sex is 121%!

Table: Summary statistics for first names sample

	Contact rate	# apps	# first names	Wald test of heterogeneity
Male				
Black	0.233	20,927	19	12.6
	(0.003)			[0.82]
White	0.246	20,975	19	15.8
	(0.003)			[0.61]
Female				
Black	0.226	20,879	19	21.2
	(0.003)			[0.24]
White	0.254	20,862	19	19.9
	(0.003)			[0.34]
Estimated contact rate SD				
Total	0.010			
Between race/sex	0.011			

Defining θ

Let N_i give # of apps sent with first name i and C_i give # of contacts within 30 days.

Assuming $C_i | N_i = n \sim Bin(n, p_i)$ we have

$$\mathbb{E}[C_i/N_i] = p_i, \quad \mathbb{V}[C_i/N_i] = p_i(1-p_i)/N_i$$

Stabilize variance with Bartlett (1936) transform

$$\hat{\theta}_i = \sin^{-1}\sqrt{C_i/N_i}.$$

Why this helps: $\frac{d}{dx}\sin^{-1}\sqrt{x} = \left[2\sqrt{x(1-x)}\right]^{-1}$. Hence, by the Delta method

$$\hat{\theta}_i \mid N_i \sim \mathcal{N}(\theta_i, (4N_i)^{-1}), \text{ where } \theta_i = \sin^{-1}(p_i).$$

Estimating G

Hierarchical model:

$$\hat{\theta}_i | \theta_i \sim \mathcal{N}(\theta_i, (4N_i)^{-1})$$

$$\theta_i | N_i \sim G(\theta)$$

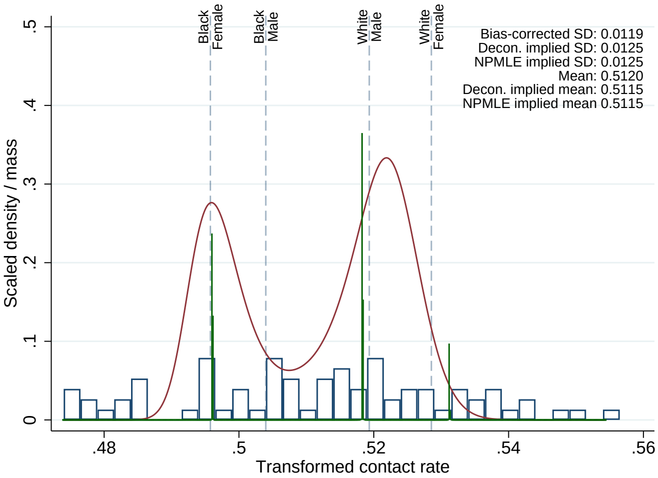
Empirical Bayes: Estimate G via deconvolution, then treat $\hat{G}$ as prior

Two approaches to deconvolution:

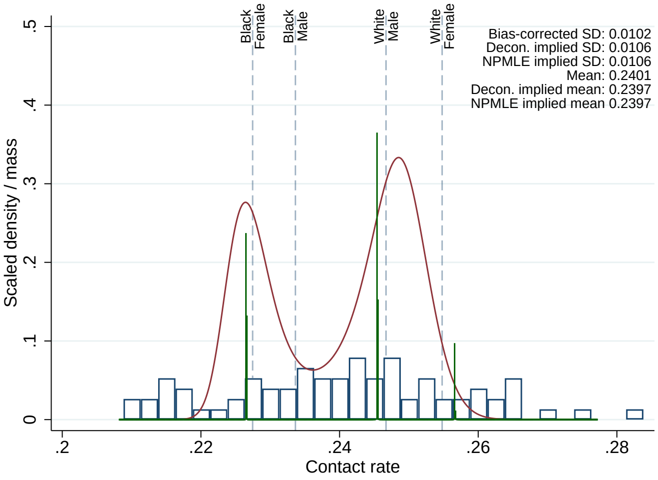
- Efron (2016): model G with exponential family parameterized by fifth-order spline, estimate via penalized MLE
- Koenker and Gu (2017): mass point approximation via NPMLE

True G seems likely to be smooth $\mapsto$ focus on Efron approach, which implies ties are measure zero

Variance-stabilized contact rates ($\sin^{-1}\sqrt{p_i}$)



Contact rates (p_i)



Empirical Bayes posteriors and grades

EB posterior density for θ_i:

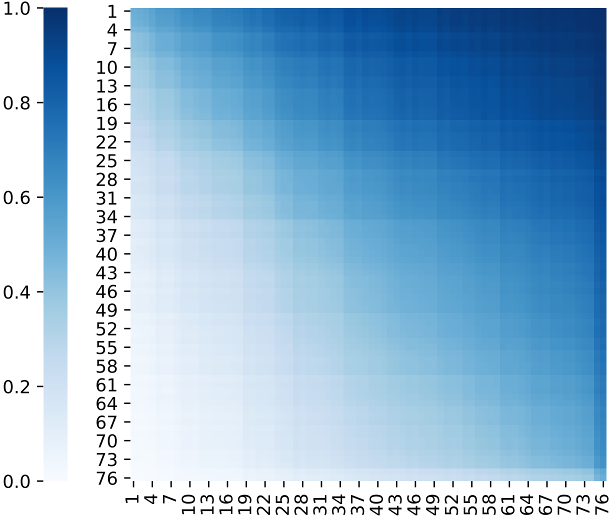
$$\hat{f}(\theta_i|\hat{\theta}_i, s_i) = \frac{\frac{1}{s_i}\phi\left(\frac{\hat{\theta}_i - \theta_i}{s_i}\right) d\hat{G}(\theta_i|s_i)}{\int \frac{1}{s_i}\phi\left(\frac{\hat{\theta}_i - x}{s_i}\right) d\hat{G}(x|s_i)}$$

Here, std err is $s_i = (4N_i)^{-1/2}$. Pairwise posterior probabilities are:

$$\hat{\pi}_{ij} = \int_{-\infty}^{\infty}\int_{-\infty}^{x} \hat{f}(x|\hat{\theta}_i, s_i)\hat{f}(y|\hat{\theta}_j, s_j)dydx$$

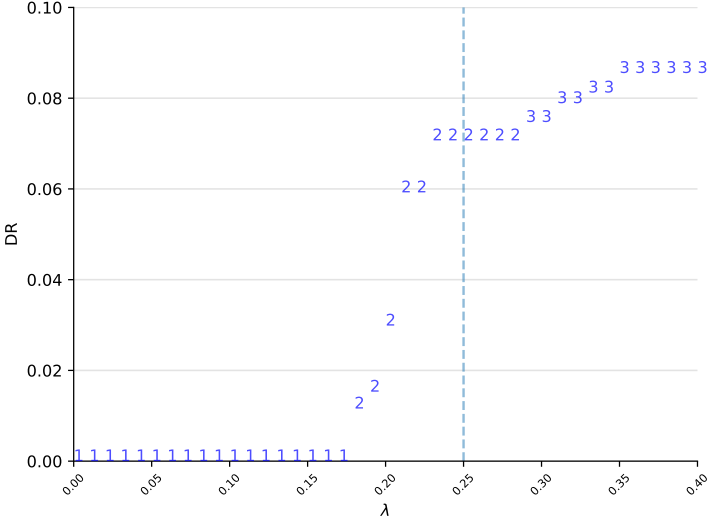
Feed these $\hat{\pi}_{ij}$'s to integer linear programming routine to compute optimal grades for each value of the tuning parameter λ

Posterior contrasts (π_{ij})

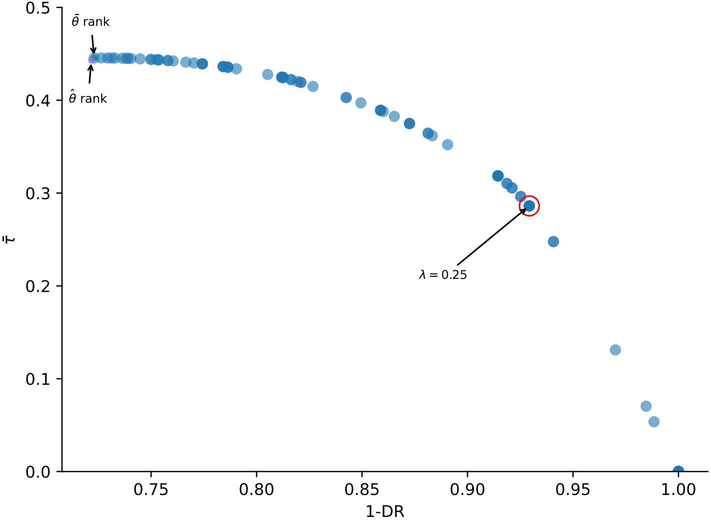


Note: Firms ordered by rank under $\lambda = 1$. Rank implying largest θ_i denoted by 1.

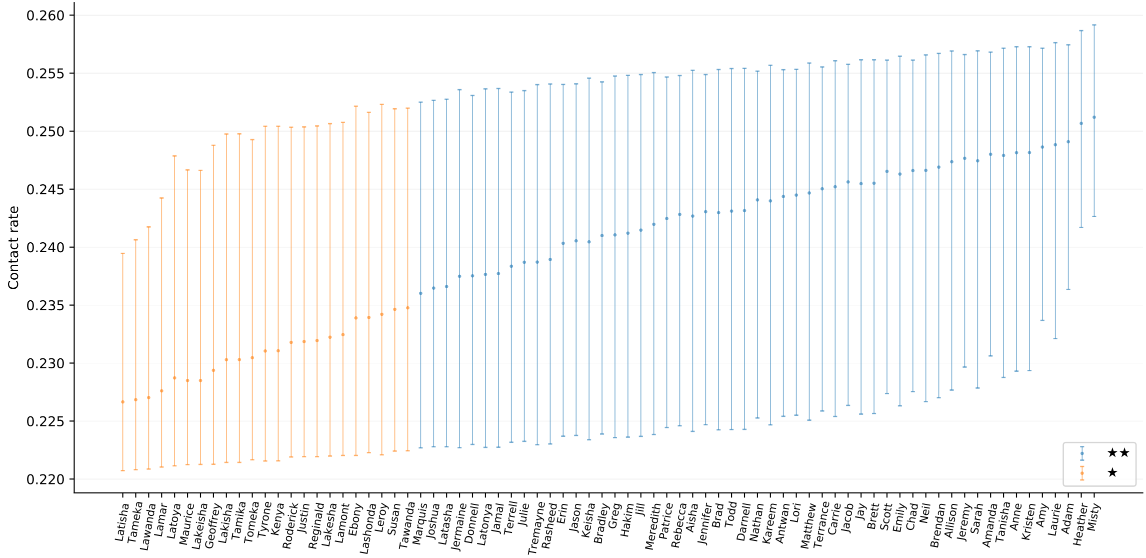
Tune grades to exhibit $\sim$ 80% posterior confidence threshold



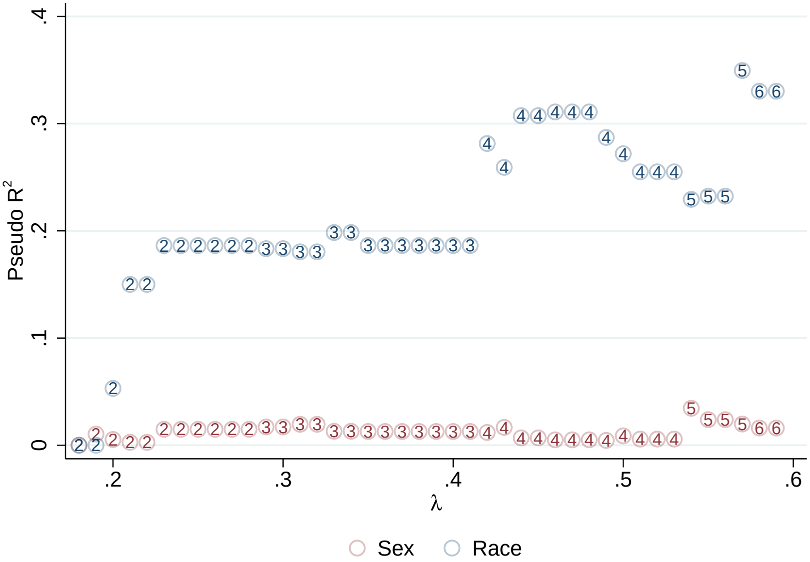
Reporting possibilities



Two grade scheme explains 35% of cross name variance



Grades predict race but not sex



Empirics: Firms

Defining θ

Each firm i has latent pair (p_{iw}, p_{ib}) of race-specific application contact rates

Focus on proportional contact gap between white and Black applicants:

$$\theta_i = \ln(p_{iw}) - \ln(p_{ib})$$

Plug-in estimator of θ_i:

$$\hat{\theta}_i = \ln(\hat{p}_{iw}) - \ln(\hat{p}_{ib}),$$

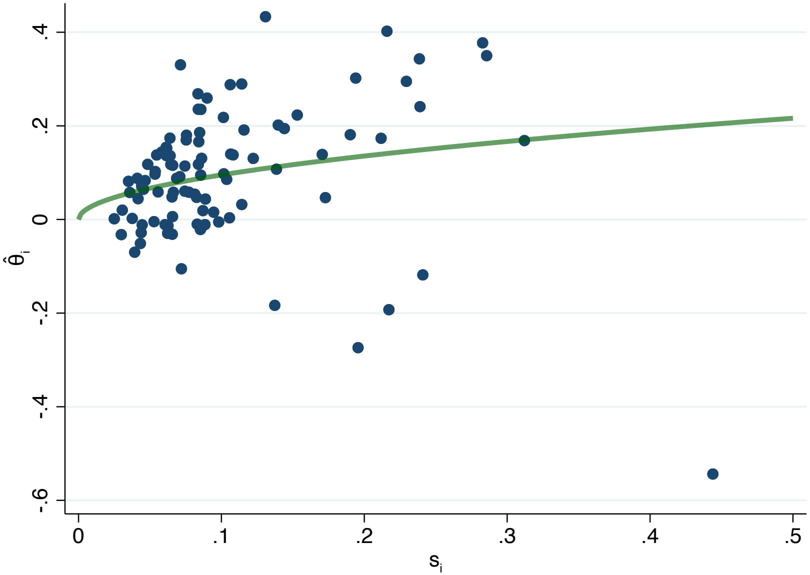
where $(\hat{p}_{ib}, \hat{p}_{iw})$ are sample averages. Standard errors $s_i = \sqrt{\hat{\mathbb{V}}[\hat{\theta}_i]}$ computed via Delta method.

Drop firms with fewer than 40 sampled jobs or callback rates $< 3\%$, leaving $n = 97$

Firm sample summary statistics

	# Firms	# Jobs	# Apps	Contact rates and gaps				
				White	Black	Difference	Log dif	Mean SE
All	97	10,453	78,910	0.256 (0.016)	0.236 (0.016)	0.020 (0.003)	0.095 (0.015)	0.095
2-digit SIC industry (code)								
Food products (20)	1	100	788	0.435 (0.000)	0.440 (0.000)	-0.005 (0.000)	-0.011 (0.000)	0.045
Apparel manufacturing (23)	2	200	1,538	0.205 (0.028)	0.175 (0.037)	0.031 (0.010)	0.177 (0.082)	0.088
Other manufacturing (24)	4	375	2,904	0.119 (0.037)	0.104 (0.037)	0.015 (0.003)	0.179 (0.061)	0.211
Freight / transport (42)	4	458	3,300	0.194 (0.019)	0.197 (0.020)	-0.003 (0.002)	-0.014 (0.011)	0.076
Communications (48)	2	175	1,124	0.273 (0.147)	0.225 (0.113)	0.048 (0.035)	0.163 (0.055)	0.120
Electric / gas (49)	3	320	2,419	0.261 (0.100)	0.247 (0.112)	0.014 (0.014)	0.120 (0.076)	0.094
Wholesale durable (50)	2	152	1,143	0.194 (0.035)	0.177 (0.027)	0.017 (0.008)	0.088 (0.030)	0.081
Wholesale nondurable (51)	11	1,117	8,194	0.299 (0.066)	0.288 (0.073)	0.011 (0.009)	0.092 (0.035)	0.091
Building materials (52)	3	377	2,755	0.297 (0.125)	0.285 (0.116)	0.012 (0.013)	0.024 (0.029)	0.062
General merchandise (53)	12	1,380	10,440	0.320 (0.048)	0.292 (0.045)	0.028 (0.006)	0.108 (0.029)	0.083
Food stores (54)	5	530	4,030	0.451 (0.089)	0.425 (0.086)	0.026 (0.010)	0.063 (0.031)	0.058
Auto dealers / services (55)	8	891	6,930	0.257 (0.041)	0.204 (0.034)	0.053 (0.011)	0.237 (0.042)	0.107
Apparel stores (56)	4	400	3,093	0.237 (0.075)	0.202 (0.064)	0.035 (0.015)	0.173 (0.057)	0.117
Furnishing stores (57)	4	482	3,679	0.286 (0.037)	0.251 (0.033)	0.035 (0.004)	0.131 (0.005)	0.086
Eating/drinking (58)	4	500	4,000	0.368 (0.027)	0.337 (0.020)	0.032 (0.009)	0.086 (0.021)	0.053
Other retail (59)	7	816	6,281	0.206 (0.056)	0.182 (0.048)	0.024 (0.010)	0.133 (0.057)	0.138
Banks / credit (60)	2	252	1,947	0.119 (0.058)	0.121 (0.048)	-0.002 (0.011)	-0.073 (0.120)	0.150
Securities brokers (62)	1	125	965	0.122 (0.000)	0.111 (0.000)	0.011 (0.000)	0.098 (0.000)	0.102
Insurance / real estate (63)	5	398	2,907	0.142 (0.074)	0.142 (0.076)	0.000 (0.010)	0.015 (0.157)	0.203
Accommodation (70)	2	243	1,850	0.200 (0.037)	0.199 (0.064)	0.001 (0.027)	0.043 (0.148)	0.094
Business services (73)	3	375	2,812	0.214 (0.118)	0.212 (0.127)	0.003 (0.010)	0.101 (0.102)	0.113
Auto / repair services (75)	3	340	2,551	0.285 (0.021)	0.275 (0.032)	0.010 (0.014)	0.046 (0.053)	0.062
Health services (80)	4	400	2,886	0.150 (0.062)	0.144 (0.048)	0.006 (0.017)	-0.071 (0.101)	0.127
Engineering services (87)	1	47	374	0.122 (0.000)	0.117 (0.000)	0.005 (0.000)	0.044 (0.000)	0.042

Standard errors predict point estimates



A model of precision-dependence

Work with a model of proportional dependence:

$$\theta_i = s_i^{\beta} v_i,\ v_i | s_i \sim G_v$$

- Assume $G_v(0) = 0$: no firm prefers Black names (test yields $p = 0.94$).
- Estimate β along with $\mu_v \equiv \mathbb{E}[v_i]$ and $\sigma_v^2 \equiv \mathbb{V}[v_i]$ via GMM
- Deconvolve standardized residual $\hat{v}_i = \hat{\theta}_i / s_i^{\hat{\beta}}$ ala Efron (2016) to recover $\hat{G}_v$
- Choose logspline tuning parameter to match GMM estimates of μ_v and σ_v^2

Building in industry effects

Allow random effect for industry $k(i)$:

$$\hat{\theta}_i = s_i^{\beta} \times \underbrace{\eta_{k(i)}}_{\text{Industry effect}} \times \underbrace{\xi_i}_{\text{Firm Effect}} + \underbrace{s_i \times \epsilon_i}_{\text{Noise}},$$

$$\epsilon_i | s_i, \eta_{k(i)}, \xi_i \sim N(0,1)$$

$$\xi_i \mid s_i, \eta_{k(i)} \sim G_{\xi},$$

$$\eta_k \mid \mathbf{s}_k \sim G_{\eta},$$

$$\mathbb{E}[\xi_i] = \mu_v, \quad \mathbb{E}[\eta_k] = 1.$$

- Extend Efron (2016)'s deconvolution estimator to hierarchical case, modeling G_{ξ} and G_{η} with two fifth-order splines with non-negative support.
- Form posteriors for each θ_i given estimates $\hat{G}_{\eta}$ and $\hat{G}_{\xi}$ along with estimates $\{\hat{\theta}_j, s_j\}_{j:k(j)=k(i)}$ for all firms in the same industry

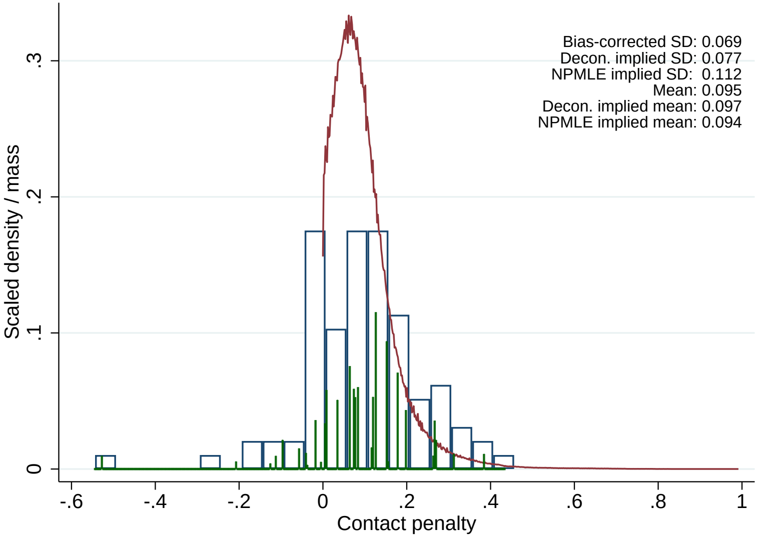
Results

Table: GMM Estimates of Contact Penalty Parameters

	No industry effects (1)	With industry effects (2)
β	0.510	0.517
	(0.190)	(0.121)
μ_v	0.313	0.292
	(0.074)	(0.074)
σ_v	0.207	
	(0.106)	
σ_η		0.452
		(0.171)
σ_ξ		0.144
		(0.066)
Within share		0.556
J-statistic (d.f.)	0.101 (1)	0.111 (2)

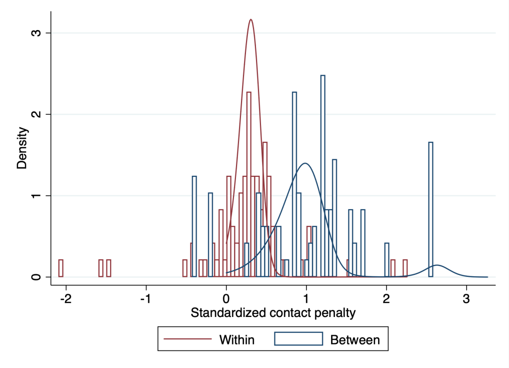
Implications: $\theta_i \approx \sqrt{s_i} v_i$ and roughly $1/2$ of variance of v_i within industry.

Deconvolution estimates reveal substantial heterogeneity in conduct

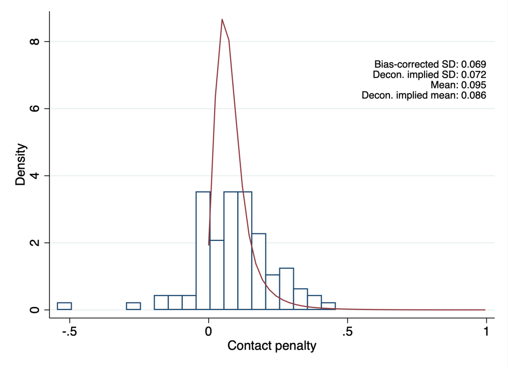


Significant variation within and between industries

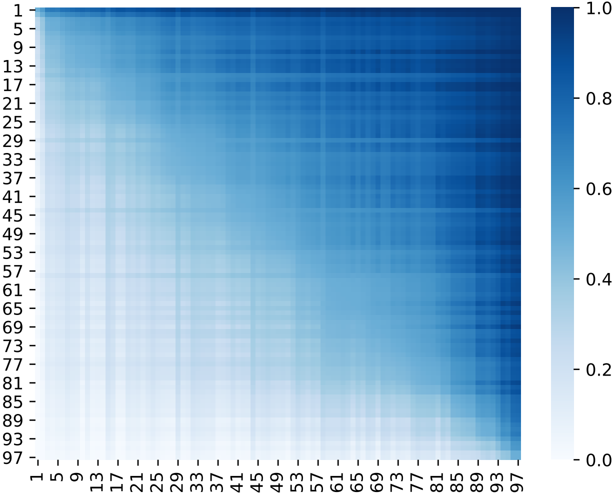
a) Within- and between-industry components



b) Marginal distribution of θ_i

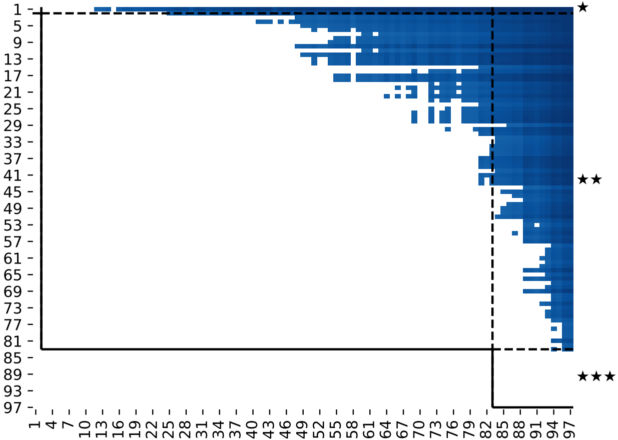


Posterior contrasts (π_{ij})



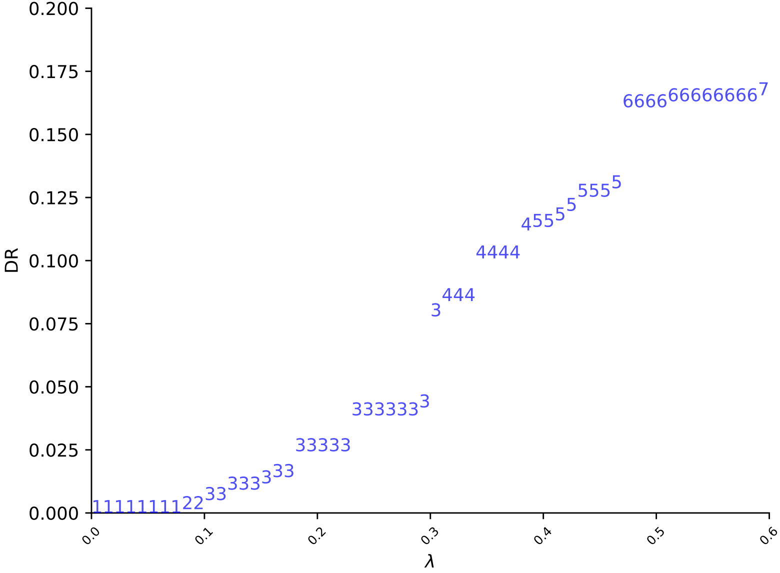
Note: Firms ordered by rank under $\lambda = 1$. Rank implying largest θ_i denoted by 1.

Pairwise decisions and optimal grades when $\lambda = 0.25$

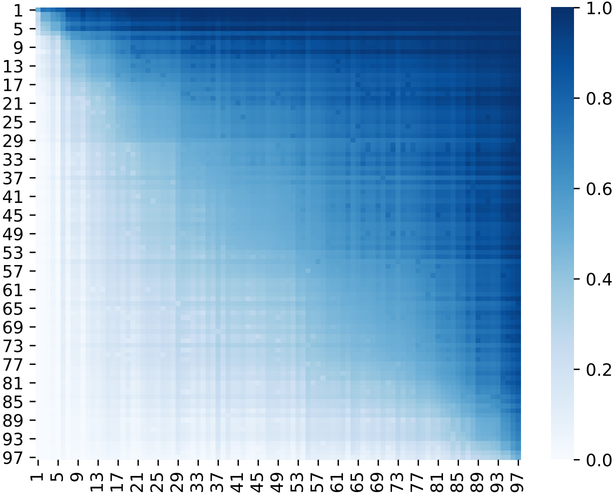


Note: Firms ordered by rank under $\lambda = 1$. Rank implying largest θ_i denoted by 1.

Discordance Rate and # of grades by λ

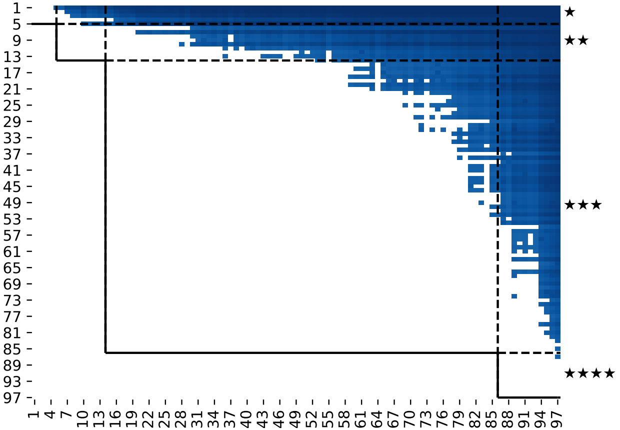


Posterior contrasts and grades with industry effects



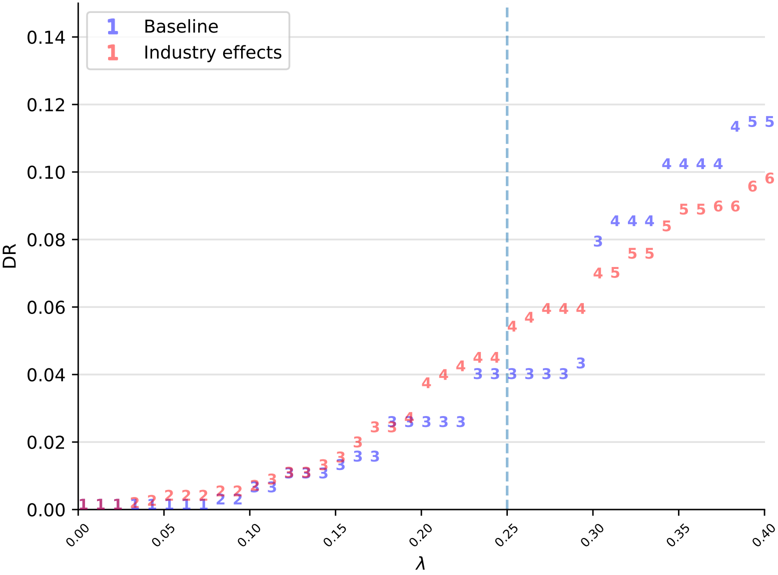
Note: Firms ordered by rank under $\lambda = 1$. Rank implying largest θ_i denoted by 1.

Posterior contrasts and grades with industry effects

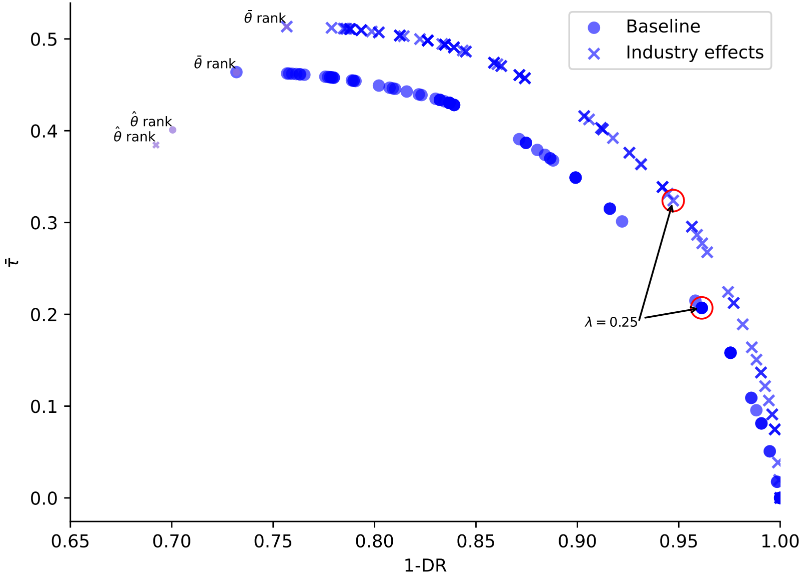


Note: Firms ordered by rank under $\lambda = 1$. Rank implying largest θ_i denoted by 1.

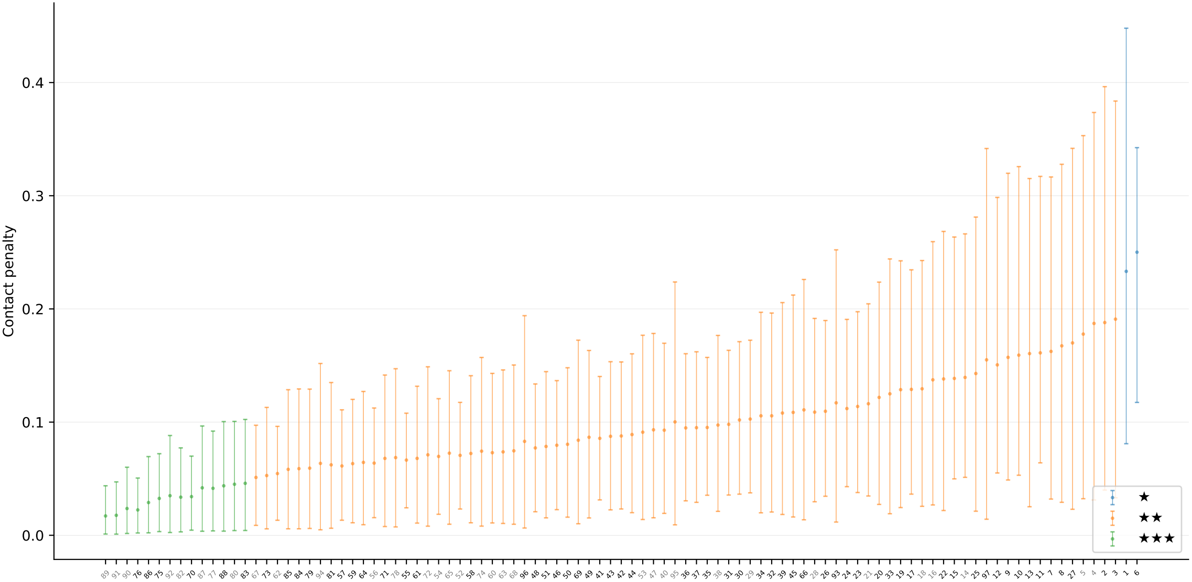
Discordance Rate and # of grades by λ



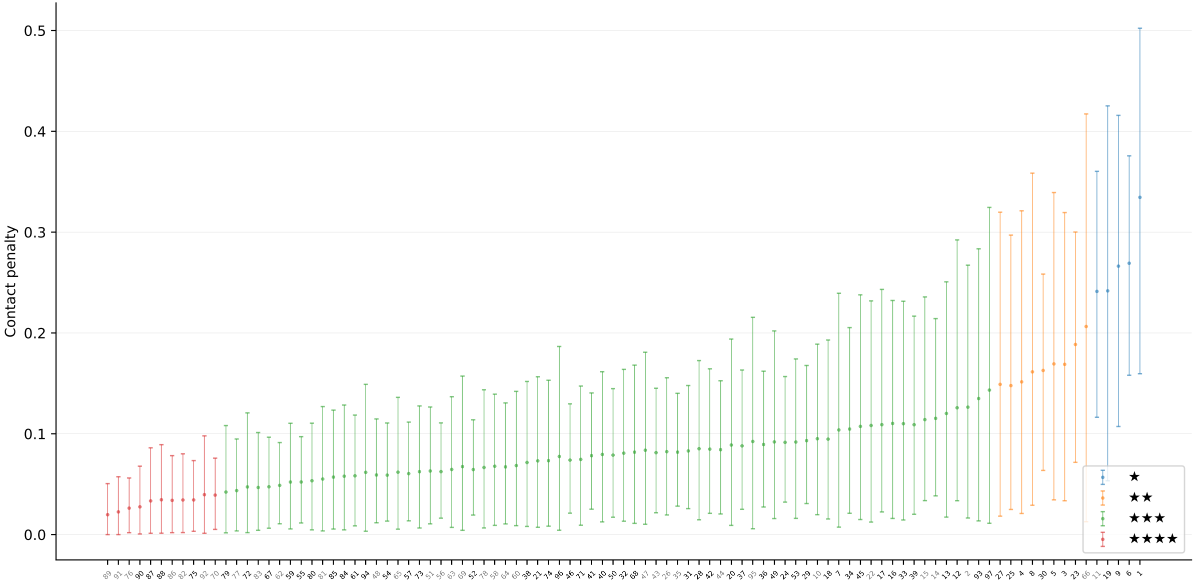
Reporting possibilities



Posterior distributions and grades, no industry effects



Posterior distributions and grades, with industry effects



Some observations

Top 4 discriminators are fed contractors subject to OFCCP oversight

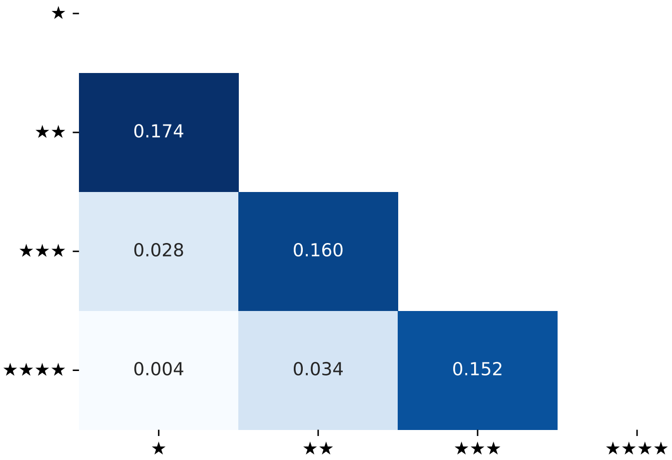
- Fed contractors less biased *on average* but comprise 2/3rds of our sample.
- Top 4 exhibit posteriors means $> 20\%$
- Potential violation of "4/5ths rule" from Uniform Guidelines (1978)

  A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact.

Accepting vs failing to reject a null

- Average posterior bias among firms graded as $\star$: 22%
- Average posterior bias among firms graded as $\star\star\star\star$: 3%

Rare to misclassify by more than one grade



Conclusion

New approach to ordinal reporting when concerned about misclassification

- Simple idea: maximize $\mathbb{E}_G[\tau(\theta, d)|Y]$ while limiting DR
- Applicable to many other reporting tasks involving value added or conduct

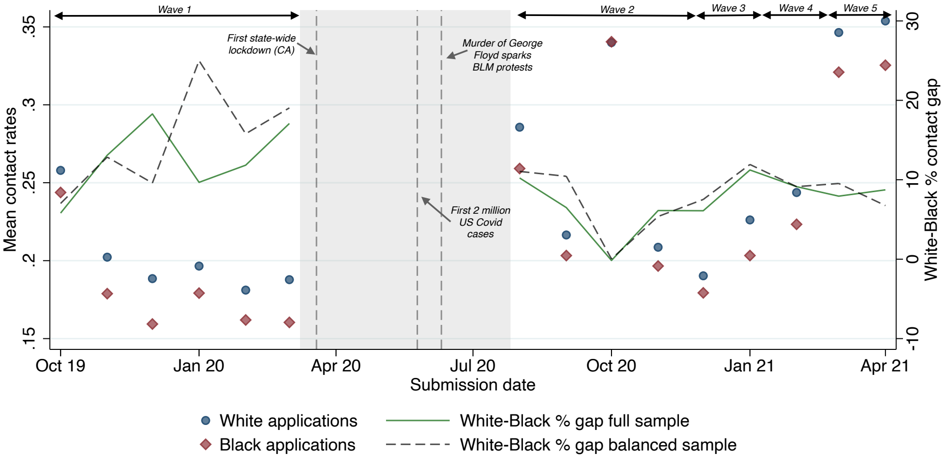
How much information about discriminatory conduct can be reliably communicated?

- With n grades: $\tau = 0.46$, $DR = 0.27$ (or $\tau = 0.51$, $DR = 0.24$ w/ industry effects)
- Fixing $\lambda = 0.25$ yields 3 grades, $\tau = 0.21$, and $DR = 0.04$ (or 4 grades, $\tau = 0.32$, $DR = 0.05$ w/ industry effects)

Ranking package `DRrank` available at `https://github.com/ekrose/drrank`

- Works with any set of posterior probs π_{ij}
- Rapid computation for $n < 500$

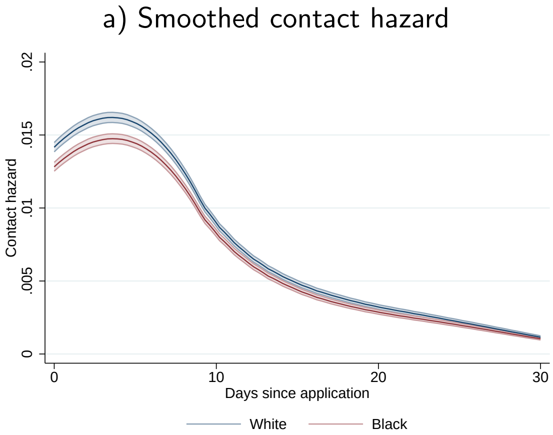
Bonus material

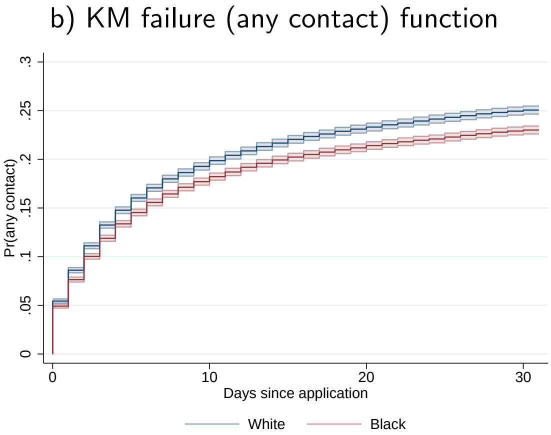


Average Black/white contact gap of 2.1pp, or 9%

- 36% avg. gap reported in meta-analysis of Quillian et al. (2017)
- Level diffs of 3pp in Bertrand and Mullainathan (2004) and 2.6pp in Nunley et al. (2015)
- Discrimination less severe among large firms? (Banerjee et al. 2018)

Contact gap stabilizes by 30 days





Extension: weighted loss

Large mistakes more costly. Consider augmented loss function $L^p(\theta, d; \lambda) =$

$$\binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i} \Bigg[\underbrace{1\{\theta_i > \theta_j, d_i < d_j\}(\theta_i - \theta_j)^p + 1\{\theta_i < \theta_j, d_i > d_j\}(\theta_j - \theta_i)^p}_{\text{discordant pairs}} -$$

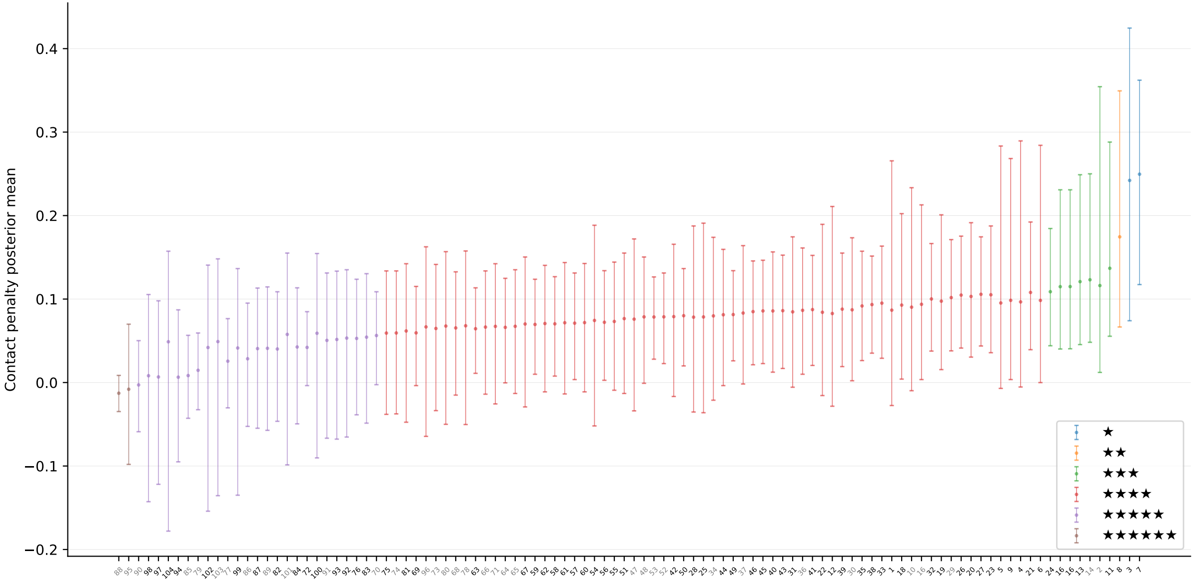
$$\lambda \Bigg(\underbrace{1\{\theta_i < \theta_j, d_i < d_j\}(\theta_i - \theta_j)^p + 1\{\theta_i > \theta_j, d_i > d_j\}(\theta_j - \theta_i)^p}_{\text{concordant pairs}} \Bigg) \Bigg].$$

The corresponding Bayes risk function takes the linear form

$$\binom{n}{2}^{-1} \sum_{i=2}^{n} \sum_{j=1}^{i} \mu_{ji}^p d_{ij} + \mu_{ij}^p (1 - e_{ij} - d_{ij}) - \lambda \mu_{ji}^p (1 - e_{ij} - d_{ij}) - \lambda \mu_{ij}^p d_{ij},$$

where $\mu_{ij}^p = \mathbb{E}_G[\max\{(\theta_i - \theta_j), 0\}^p \mid Y_i = y_i, Y_j = y_j]$.

Square-weighted loss: Posterior means and grades (baseline)



Square-weighted loss: Posterior means and grades (industry FX)