NAMESAKES ON THE LIST: COMPOSITION AMPLIFICATION AND TARGET-DECOY CALIBRATION IN SANCTIONS NAME SCREENING

Anonymous authors
Paper under double-blind review

ABSTRACT

Sanctions screening compares each customer's name with every entry on a watchlist of tens of thousands of listed persons and their aliases, and raises an alert when any entry scores above a similarity threshold. We show that this one-to-many design, more than the pairwise matcher, decides who bears the false alerts. Screening non-sanctioned politically exposed persons from 36 countries against a consolidated list of about 40,000 sanctioned persons, a standard fuzzy matcher at a common operating threshold alerts on people from the countries most represented on the list more than an order of magnitude more often than on others. We formalize this composition amplification: a customer's false-alert probability grows with the number of listed namesakes, so no single score threshold can equalize alert burden across populations whose names collide with the list at different rates. We then adapt target-decoy competition from proteomics: each customer is screened against the real list and against decoy lists of non-sanctioned persons drawn to match the list's composition, and is flagged only when the real list wins. When listed and decoy names are exchangeable, the resulting per-customer p-value bounds every customer's false-alert probability, and hence every group's, without using the customer's demographic attributes. On real lists the guarantee is only as good as the decoys: we show how composition misspecification shifts false alerts between groups rather than removing them, and quantify the recall the calibration costs for listed persons with common names. We evaluate string, embedding and large-language-model matchers and release the benchmark.

1 INTRODUCTION

Banks, payment processors and customs administrations must compare the name of every customer and counterparty against consolidated sanctions lists, and have a human review whatever the comparison flags. The comparison is one-to-many: one recorded name is scored against tens of thousands of listed names and their aliases, and an alert is raised when the best score clears a threshold. Almost all of those alerts are false; the question this paper makes precise is not how many a system produces but *whose* names produce them.

What decides a customer's fate is not the matcher's pairwise accuracy but their *namesake count*: how many entries of the list the matcher cannot separate from their name. Because a watchlist is a highly non-uniform sample of the world's name populations, namesake counts are structurally unequal. Screening 40,755 non-sanctioned politically exposed persons from 36 countries against a consolidated list of 39,678 sanctioned persons with a standard fuzzy matcher at a common operating threshold, we find alert rates of 38.8% for the most represented country and 0.13% for the least, a ratio of 291 to one; the five most affected countries are 75 times more likely to be alerted on than the five least affected. The same pattern appears for every matcher family we tried, including neural and LLM-based matchers, because it is a property of the one-to-many design and of the list, not of the scoring function. The mechanism is known in one-to-many face identification, where per-group false-positive rates grow with the group's presence in the gallery (National Institute of Standards and Technology, 2024; Howard et al., 2019); we show that it governs name screening, formalise it as a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

namesake-count law (Proposition 1), and show that a simple composition model fitted on *non-listed* names predicts the observed per-country alert rates (Spearman $\rho = 0.88$).

The consequence is uncomfortable: no single threshold can equalise alert burden, and the obvious repair, a threshold per name population, requires the screening system to condition on something close to ethnicity. We therefore ask whether the burden can be equalised *without* the attribute. Our answer adapts target-decoy competition, the device mass spectrometry uses for exactly this problem: a spectrum's best database hit is meaningless until one knows how crowded the database is around it, so the search is repeated against decoys generated from the database itself and a hit is kept only when the target wins (Elias & Gygi, 2007). Screening each customer against the real list and against K decoy lists of non-listed names drawn to match the list's composition, and flagging only when the real list wins, yields a rank-based p-value whose validity is conditional on the individual customer and therefore holds simultaneously for every group, observed or not (Proposition 2).

The guarantee rests on one assumption, that the real list and the decoy lists are exchangeable, and the substance of the paper is how much that assumption costs on real data. We prove that the error is bounded by a total-variation distance between two screening-score laws (Proposition 3), that misspecification *moves* false alerts between groups rather than removing them (Corollary 4), and that the same namesake density that drives the burden also caps detection: a listed person whose name is common in their own population cannot be flagged by any calibrated name-only rule (Proposition 5). Empirically, when decoys are exchangeable by construction the worst country's false-alert rate sits at 1.17α, inside the band that binomial noise alone produces, while every decoy source we can build from public data lands at 3.21α to 4.43α, far outside it. The failure is measurable in advance: it tracks a decoy-supply statistic that we can compute without labels, and it disappears as the supply ratio grows.

Contributions.

- **Composition amplification.** A namesake-count law for one-to-many name screening (Proposition 1), measured on a real sanctions list and a real non-sanctioned population across 36 countries and six matchers, with an out-of-sample predictor of per-country burden built only from non-listed names and the list's composition.
- **Target-decoy calibration for screening.** A demographic-blind rule with a per-customer, hence per-group, false-alert guarantee (Proposition 2), a total-variation robustness bound (Proposition 3), and an identifiability limit for the recall it costs (Proposition 5).
- **What decoys must satisfy.** An anatomy of decoy design (composition, alias multiplicity, supply), a validation protocol in which exchangeability holds by construction, and a label-free diagnostic that predicts calibration failure before deployment.
- **Benchmark.** A single-machine, fully reproducible benchmark from dated public snapshots, so that a decoy model can be scored the way a matcher is scored today.

2 SETTING: ONE-TO-MANY SCREENING AND COMPOSITION AMPLIFICATION

Screening. Let $\mathcal{X}$ be a set of name strings and $s : \mathcal{X} \times \mathcal{X} \to [0, 1]$ a matcher that scores a pair of names. A *watchlist* L is a finite multiset of name strings: each listed person contributes a primary name and any aliases the list records for them. A customer with recorded name x is screened by comparing x with *every* entry of the list and keeping the best score,

$$S(x, L) \;=\; \max_{y \in L} s(x, y), \qquad A_\tau(x, L) \;=\; \mathbf{1}\{S(x, L) \geq \tau\}, \tag{1}$$

where A_τ is the alert that a human analyst must then adjudicate. We call $N_\tau(x, L) = |\{y \in L : s(x, y) \geq \tau\}|$ the *namesake count* of x: the number of list entries that the matcher cannot separate from x at threshold τ. An alert is raised exactly when $N_\tau \geq 1$.

Every customer carries an attribute $g(x) \in \mathcal{G}$ that indexes the name population they are drawn from (in our experiments, the country of the customer). The screening system never observes g; it is used only to measure who bears the alerts. Write $\mathrm{FAR}_g(\tau) = \mathbb{P}(A_\tau = 1 \mid g(x) = g,\ H_0(x))$ for the *false-alert rate* of group g, where the null $H_0(x)$ states that no entry of L refers to x.

A random-list model. List entries are names of real people, so they inherit the name distribution of the populations the list draws from. Let F_g be the distribution of name strings in population g and let $\pi = (\pi_g)_{g\in\mathcal{G}}$ be the *composition* of the list: the share of its entries drawn from each population. Under model $\mathcal{M}(n, \pi, F)$ the list is $L = \{y_1, \dots, y_n\}$ with y_i drawn independently from $F_\pi = \sum_g \pi_g F_g$. Define the per-entry collision probability and the expected namesake count

$$q_g(x,\tau) = \mathbb{P}_{y\sim F_g}[s(x,y) \geq \tau], \qquad \Lambda(x,\tau) = n\sum_{g\in\mathcal{G}} \pi_g\, q_g(x,\tau). \tag{2}$$

Proposition 1 (Composition amplification)**.** *Under $\mathcal{M}(n, \pi, F)$, for any customer x satisfying $H_0(x)$,*

$$\mathbb{P}[A_\tau(x,L)=1] = 1-\Big(1-\tfrac{\Lambda(x,\tau)}{n}\Big)^n, \qquad \textit{so} \qquad 1-e^{-\Lambda(x,\tau)} \;\leq\; \mathbb{P}[A_\tau=1] \;\leq\; \min\{1,\Lambda(x,\tau)\}, \tag{3}$$

and $\mathbb{E}\, N_\tau(x,L) = \Lambda(x,\tau)$. Consequently the alert probability of a customer is a strictly increasing function of their expected namesake count, and this is the only channel through which the list affects them. Moreover:

- *(i)* (Share, not skill.) *If collisions are within-population, i.e. $q_h(x,\tau) \approx 0$ for $h \neq g(x)$, then $\Lambda(x,\tau) \approx n\,\pi_{g(x)} q_{g(x)}(x,\tau)$. If in addition the matcher is* pairwise calibrated *across groups at τ, meaning $\bar{q}_g(\tau) := \mathbb{E}_{x\sim P_g} q_g(x,\tau)$ does not depend on g, then for small Λ the burden ratio between two groups is $\mathrm{FAR}_a(\tau)/\mathrm{FAR}_b(\tau) \to \pi_a/\pi_b$: it is set by the list's composition and not by the matcher.*
- *(ii)* (No single threshold equalises.) *If, for every τ in an operating range T, the expected namesake count $\Lambda(\cdot,\tau)$ under P_a stochastically dominates that under P_b, then $\mathrm{FAR}_a(\tau) \geq \mathrm{FAR}_b(\tau)$ for every $\tau \in T$, with strict inequality whenever the dominance is strict. Equal burden can then be obtained only by thresholds $\tau_a \neq \tau_b$ that condition on the customer's group.*

Proposition 1 is the name-screening version of a law documented for one-to-many face identification, where a demographic's false-positive identification rate grows with its presence in the gallery (National Institute of Standards and Technology, 2024; Howard et al., 2019; Grother et al., 2019; Sirotin & Vemury, 2021). Its force here comes from how concentrated sanctions lists are: the largest country stratum holds 41% of all name strings while the smallest customer population we screen contributes under 0.1%, so part (i) predicts burden ratios in the hundreds before any matcher is chosen, and part (ii) says a compliance function that wants equal burden must either accept unequal burden or screen people differently by an attribute it does not hold and may not lawfully use.

3 Target-decoy calibration for screening

The procedure. Proteomics faces a structurally identical problem: a spectrum is matched against every peptide in a database and the best hit is reported, so the significance of a top score depends on how crowded the database is around that spectrum. The standard answer is *target-decoy competition*: the search is repeated against a decoy database generated from the target database itself, and a match is kept only when the target wins (Elias & Gygi, 2007; 2010; Keich et al., 2015). We transfer the construction to screening.

Let $D_1, \dots, D_K$ be *decoy lists*: multisets of names of persons who are not on the watchlist, built to imitate the watchlist (Section 3.1). Screen the customer against the real list and against every decoy list, $S_0 = S(x, L)$ and $S_j = S(x, D_j)$, and report the rank-based p-value

$$p(x) \;=\; \frac{1 + |\{j \in [K] : S_j \geq S_0\}|}{K+1}. \tag{4}$$

The screening rule is: alert if $p(x) \leq \alpha$ (we also require $S_0 \geq \tau_{\min}$, a floor below which no analyst would look; this only makes the rule more conservative). The customer is flagged when the real list beats the decoys, not when the raw score is high.

Assumption 1 (Exchangeable decoys)**.** *Conditional on the customer's name x and on $H_0(x)$, the lists $(L, D_1, \dots, D_K)$ are exchangeable.*

Proposition 2 (Per-customer and per-group validity)**.** *Under Assumption 1, $p(x)$ is super-uniform conditional on x: $\mathbb{P}[p(x) \le \alpha \mid x] \le \alpha$ for every $\alpha \in [0,1]$. Consequently, for* every *group $G \subseteq \mathcal{X}$, measurable with respect to the customer's name or any attribute of the customer,*

$$\mathbb{P}\big[\textit{alert} \mid x \in G,\ H_0\big] \ \le\ \alpha, \tag{5}$$

and the rule attains this without observing the group.

The proof (Appendix A) is the classical rank argument for Monte-Carlo p-values (Besag & Clifford, 1989; Phipson & Smyth, 2010), applied with an unusual choice of what is resampled: *the list*, not the customer. That choice buys a guarantee conditional on the individual and therefore on every subpopulation at once, which group-wise thresholding or score calibration delivers only for groups named in advance and observed at run time (Moslemi & Milani, 2024; Hébert-Johnson et al., 2018), and it is what makes the rule demographic-blind: nothing in (4) refers to $g(x)$.

When decoys are wrong. Assumption 1 is a modelling claim about the decoy generator, and it is the whole difficulty. The following bound is the screening analogue of coverage results for conformal prediction beyond exchangeability (Barber et al., 2023).

Proposition 3 (Misspecified decoys)**.** *Suppose $D_1, \dots, D_K$ are i.i.d. draws from a decoy distribution Q and the real list is drawn from P, independently, conditionally on x. Then*

$$\Big|\mathbb{P}[p(x) \le \alpha \mid x] - \mathbb{P}_{\mathrm{exch}}[p(x) \le \alpha \mid x]\Big| \ \le\ d_{\mathrm{TV}}\Big(\mathcal{L}_P\big(S(x,\cdot)\big),\ \mathcal{L}_Q\big(S(x,\cdot)\big)\Big), \tag{6}$$

so $\mathbb{P}[p(x) \le \alpha \mid x] \le \alpha + d_{\mathrm{TV}}$. Only the law of the screening score *has to be close, not the law of the names. Moreover, if $\mathcal{L}_P(S)$ stochastically dominates $\mathcal{L}_Q(S)$ then $\mathbb{P}[p \le \alpha] \ge \mathbb{P}_{\mathrm{exch}}[p \le \alpha]$, and the inequality reverses under the opposite dominance.*

Corollary 4 (Misspecification moves the burden)**.** *Fix the alert budget: choose α so that the overall alert volume equals a target B. Groups whose names the decoys under-cover (real list stochastically closer to them than the decoys are) receive more than B, and groups the decoys over-cover receive less. Composition error therefore* redistributes *false alerts rather than removing them.*

What calibration costs. Calibration is not free, and the price is paid by exactly the people whose names are common in their own population.

Proposition 5 (Namesake identifiability limit)**.** *Let x be the recorded spelling of a listed person whose own list entry scores $\sigma = s(x, y^\star)$, and let $\rho(\sigma) = \mathbb{P}[S(x,D) \ge \sigma]$ be the probability that one decoy list also reaches σ. Then the detection probability of the rule at level α is $\mathbb{P}[p(x) \le \alpha] = \mathbb{P}\big[\mathrm{Bin}(K, \rho(\sigma)) \le \alpha(K+1) - 1\big]$, which tends to $\mathbf{1}\{\rho(\sigma) < \alpha\}$ as $K \to \infty$. With $\rho(\sigma) \approx 1 - e^{-\Lambda_D(x,\sigma)}$, detection at level α requires $\Lambda_D(x,\sigma) \lesssim \alpha$: the expected number of namesakes that a composition-matched population of non-listed people produces at the customer's own match level must itself be below α.*

A name common in its own population therefore carries less than α worth of evidence, whatever the matcher, and has to be resolved with secondary identifiers rather than by lowering the bar for everyone. The same Λ that drives the burden in Proposition 1 caps detection, which is why the two cannot be traded off by tuning a threshold.

3.1 Building decoys

Assumption 1 asks for a sample of *non-listed* names drawn as the list's names were drawn. Three properties matter, and we vary them one at a time in the experiments. **(a) Composition.** The decoy list must reproduce the list's mixture over name populations; we stratify by the country recorded for each listed person, and the 22% of listed persons who carry no country form a stratum of their own, whose decoys are drawn from donors that also carry no country. **(b) Multiplicity.** Sanctions lists record several spellings per person (3.7 name strings per listed person in our snapshot) while public registers of non-listed people record about one (1.16), so a decoy list matched on persons is a far thinner net than the real list. **(c) Supply.** A decoy list of the same size as a stratum of the real list can only be drawn from a pool of non-listed names that is substantially larger than that stratum; otherwise every decoy list is nearly the same set and the competition in (4) degenerates. We compare

four sources: PEP (non-listed politically exposed persons, person-matched), PEP-UNIF (the same pool ignoring composition), PEP+VAR (PEP donors with synthetic spelling variants matching the list's multiplicity), and LIST-DERIVED decoys generated from the list's own per-stratum token distributions with alias structure fitted to the list, dropping any generated name that collides with a listed one. The last is the analogue of the reversed decoy databases of proteomics and the only one whose supply is unlimited; a fifth, POOL-SYNTHETIC, applies the same generator to the pep pool. Appendix E gives the details.

4 BENCHMARK

Data. We build everything from two dated, publicly redistributable snapshots of OpenSanctions: the consolidated *sanctions* collection (39,678 persons after normalisation, 146,792 name strings including aliases, 169 country strata) and the *politically exposed persons* collection (704,072 persons, 815,196 strings) as of 19 and 18 September 2026. Exact dataset versions and file checksums are in Appendix B. Names are romanised with `unidecode`, lower-cased, stripped to letters, and kept when at least two tokens of length ≥ 2 remain.

Populations. The *list* is every person in the sanctions collection. The *customers* are 40,755 politically exposed persons who are *not* on the list, sampled from 36 countries (at most $1{,}500$ per country and never more than half of a country's persons); the remaining 663,317 non-listed persons form the decoy pool, which is disjoint from the customer sample. Customers stand in for a bank's book: they are real, current, non-sanctioned names with a known country, which is what the burden analysis needs. Because a customer who is secretly the listed person would count as a false alert, we check identity leakage directly: only 302 customers reach an exact name match with the list, and among those with a recorded birth year on both sides the agreement rate is 5.8% against a 1.48% chance baseline, so at most a handful of the alerts we count are true matches (Appendix C).

Detection queries. We sample 8,000 listed persons with at least two recorded spellings and hold one spelling out as the *recorded* name of a customer who really is that person; the held-out string is removed from the list for that query only. Recall is the share of these queries that raise an alert.

Matchers. The primary matcher is the token-sorted indel ratio (`rapidfuzz token_sort_ratio`), the standard cheap workhorse of deployed screening. We repeat the analysis with Jaro-Winkler, token-set ratio, a phonetic matcher (Metaphone codes per token), a character n-gram TF-IDF cosine, a multilingual sentence-embedding matcher and an LLM adjudicator. All scores are mapped to $[0, 100]$, and every matcher is used in the same one-to-many wrapper (1); the neural and LLM matchers run in the deployed two-stage form, on a frozen candidate file that is released with the benchmark (Appendix F).

Metrics. For a rule we report the overall alert rate on customers, the false-alert rate FAR_c per customer country, the ratio between the largest and the median country rate, the Gini coefficient of $\{\mathrm{FAR}_c\}$, the share of countries whose rate exceeds the nominal level, and recall on the detection queries. All p-values use $K = 199$ decoy lists and the floor $\tau_{\min} = 70$.

Compute. The benchmark runs on two CPU cores. Screening scores are exact, with no blocking, because token-sorted indel similarity with an early-exit cutoff evaluates 1.4×10^8 name pairs per second, and each decoy list is an inclusion mask over a scored universe, so K decoy lists cost one scoring pass.

5 EXPERIMENTS

5.1 WHO BEARS THE FALSE ALERTS

Screening the 40,755 non-listed customers against the full list with the token-sorted matcher at the common operating point $\tau = 85$ raises alerts on 6.3% of them, but the burden is not shared: ru customers are alerted on at a rate of 38.8%, jp customers at 0.13%, a ratio of 291 to one, and the five worst-affected countries are alerted on 75 times as often as the five least affected (Figure 1a). Raising

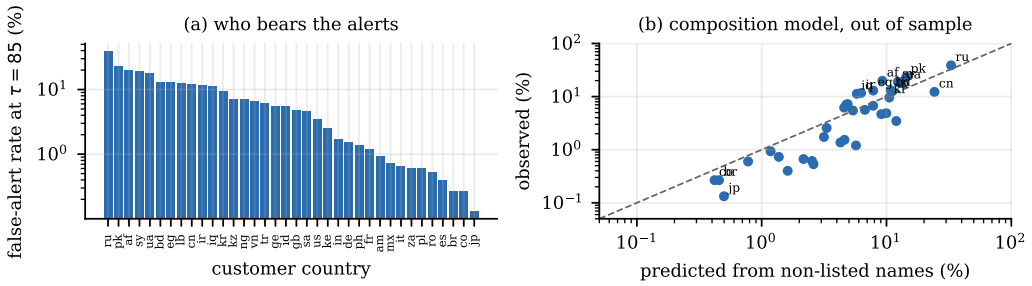

Figure 1: Composition amplification. (a) False-alert rate per customer country at a fixed threshold, log scale. (b) Rates predicted from non-listed names and the list's composition against observed rates; the dashed line is equality.

the threshold sharpens the disparity rather than removing it, exactly as Proposition 1(ii) predicts: at $\tau = 95$ the overall rate falls to 0.94% while the ratio between the largest and the median country rate grows from 7.5 to 13.1.

The mechanism is the namesake count and not the matcher. We test this out of sample: using only *non-listed* names to estimate per-stratum collision probabilities $q_g(x, \tau)$ and the list's own composition to weight them, (2) predicts each country's alert rate with Spearman correlation 0.88 and a log-scale Pearson correlation of 0.90 (Figure 1b), without ever scoring a customer against the list. The residuals are informative in both directions: countries whose listed persons carry many alias spellings (`af`, `eg`, `iq`) are under-predicted, because a list entry with eleven spellings is a wider net than one draw from the ambient distribution, and countries transliterated differently from the customer population (`cn`) are over-predicted. These are the two failure modes that later limit decoy construction. The residuals also answer the obvious worry about our customer population, that politically exposed persons might be adversarially close to a list that targets political figures: across countries the observed rates are neither systematically above nor below what non-listed names predict (median ratio 0.78, quartiles 0.45 and 1.36). Burden also does not require being on the list: `bd` customers are alerted on at 13.2% although Bangladesh contributes 35 of the 146,792 strings, because the name populations that dominate the list overlap theirs.

5.2 Is the calibration valid?

Proposition 2 holds under exchangeable decoys, so we first build a setting in which that is true by construction: listed persons are partitioned at random into B blocks, one block is screened as the list and the others act as decoys, which makes the B lists exchangeable and disjoint, as decoys must be in practice. In that setting the guarantee holds. The right yardstick is not α itself but the maximum over 36 countries of 36 noisy estimates: if every country were exactly at the nominal level, binomial fluctuation alone would put that maximum at 1.41α on average (95th percentile 1.73α) for our sample sizes. Across five random partitions at $B = 20$ and $\alpha = 0.05$ the exchangeable reference gives 1.17α (range 1.01 to 1.32), below the null band, with an overall alert rate of 0.39α: valid, and conservative because screening scores are discrete and most customers have no namesake at all. Every practical decoy source sits far outside that band: 3.21α (2.97 to 3.60) for pool-synthetic decoys, 3.38α for pep donors with matched alias structure, 3.90α for list-derived decoys and 4.43α for raw pep decoys, with 23% to 42% of countries above the nominal level. This is the reference point against which every practical decoy source can be scored.

Practical decoys never reach it. Table 1 and Figure 2 report the same experiment with four decoy sources at three list sizes. The calibration error shrinks monotonically as the list shrinks relative to the pool of non-listed names available (for list-derived decoys, from 8.0α on the full list to 2.3α at 1,601 strings), and it tracks a diagnostic computable without labels: the share of the list sitting in strata where the decoy source is exhausted. On the full list, 79% of the list's strings belong to strata in which the public register of non-listed people holds fewer than half as many names, and for person-matched sources the inclusion probability saturates for 42% of the list's strings, so the K decoy lists collapse onto nearly the same set. List-derived and pool-synthetic decoys never saturate, which is why they

Table 1: Validity as a function of list size and decoy source, $\alpha = 0.05$, $K = 199$ ($K = B - 1$ for the exchangeable reference). *max/α* is the largest country false-alert rate divided by the nominal level; recall is on detection queries whose listed person is in the screened list.

	list = 146,792 strings		7,529 strings		1,601 strings	
decoy source	max/α	recall	max/α	recall	max/α	recall
exchangeable (ideal)	–		0.9	58	0.9	63
list-derived	8.0	75	4.2	71	2.2	73
pool-synthetic	5.1	75	3.2	73	1.9	71
pep + variants	9.2	83	3.1	74	1.9	70
pep	11.3	86	4.5	74	2.0	71
pep, no composition	8.8	83	4.3	77	3.1	73

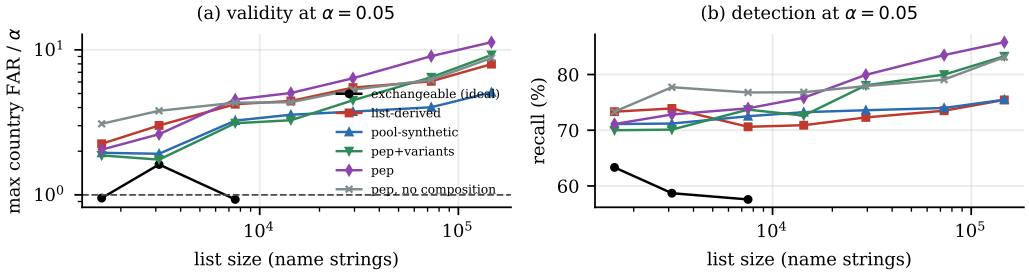

Figure 2: Calibration error and detection against list size. Exchangeable decoys stay at or below the nominal level; every practical decoy source degrades as the list grows relative to the supply of non-listed names.

degrade more gracefully, but they inherit a fidelity gap instead, and it is concentrated exactly where screening decisions are made. Real name populations repeat whole names far more often than any independent-component model does: in the validation setting 0.069% of customers have an exact match in the screened list, against 0.047% for pep donors with matched alias structure, 0.017% for list-derived and 0.002% for pool-synthetic decoys, a thirty-fold deficit at the top of the score range. A label-free repair that rescales each stratum's decoy supply until the decoys reproduce the list's own per-stratum hit counts helps in aggregate and not where it matters (Appendix G): the discrepancy is not how many namesakes the decoys produce but which customers get them.

5.3 What it buys at a fixed alert budget

Compliance functions choose a review budget, not a nominal level, so the operational question is what each rule does with the same number of alerts (Table 2). Against a threshold at the same 12.0% volume, competition with list-derived decoys cuts the Gini of country alert rates from 0.50 to 0.35, the largest country rate from 58% to 40%, and the largest-to-median ratio from 5.9 to 4.1, while detection falls only from 76.3% to 75.4%; at 5.3% volume, pool-synthetic decoys recall 75.2% against 63.9% for the threshold raising the same number of alerts. Figure 3 in Appendix D shows what the calibration does: not a uniform reduction but a redistribution, lowering the rate for the countries the list over-represents and raising it for countries whose names are rare on the list, which is what Corollary 4 describes.

A screening function does not have to choose between its threshold and the calibration: it can keep the threshold as an evidence floor and spend the same budget better. Replacing "alert if $S_0 \geq 85$" by "alert if $S_0 \geq 60$ and $p \leq 0.05$" raises 5.8% of customers instead of 6.3%, detects 75.5% of the detection queries instead of 66.3%, and brings the worst country's rate down from 38.8% to 25.5% (Gini 0.56 to 0.42). At the tighter budget the same swap turns $\tau = 90$ (2.3% volume, 52.9% recall, worst country 15.1%) into $\tau = 84$ with competition (2.2%, 61.0%, 11.2%). The threshold is doing what it is good at, excluding weak evidence, and the competition is doing what a threshold cannot, discounting evidence that is weak because the name is crowded.

Table 2: Full list, all 40,755 customers. Rules compared at matched alert volume. The oracle rules use the customer's country; the decoy rules use no attribute.

rule	volume (%)	recall (%)	max country (%)	median country (%)	max/median	Gini
threshold $\tau = 85$	6.3	66.3	38.8	5.17	7.5	0.561
threshold $\tau = 81$	12.1	76.3	58.1	9.86	5.9	0.503
per-country τ, target 1% (oracle)	0.7	13.1	1.0	0.81	1.2	0.247
per-country τ, target 5% (oracle)	4.3	26.5	5.0	4.37	1.1	0.075
TDC alone, list-derived decoys	12.0	75.4	39.8	9.80	4.1	0.351
TDC alone, pool-synthetic decoys	5.3	75.2	25.3	4.08	6.2	0.460
threshold $\tau = 80$ + TDC, list-derived	6.3	68.6	28.6	4.77	6.0	0.490
threshold $\tau = 88$ + TDC, list-derived	2.4	56.5	15.7	1.72	9.1	0.606
threshold $\tau = 60$ + TDC, pool-synthetic	5.8	75.5	25.5	5.18	4.9	0.421
threshold $\tau = 84$ + TDC, pool-synthetic	2.2	61.0	11.2	2.02	5.6	0.552

The obvious alternative is to set a threshold per country

5.4 What calibration costs, and to whom

Proposition 5 predicts a sharp transition: a listed person can be flagged only if a composition-matched population of non-listed people would produce fewer than α namesakes at their own match level. Table 4 in Appendix D confirms it on the detection queries. For the seven deciles of Λ_D whose value is zero, detection is 100%; in the top two deciles, where a decoy list contains a name as close as the person's own entry with probability 0.1 or more, detection at $\alpha = 0.01$ is zero. In total 21% of listed persons in our detection set fall above the limit. A fixed threshold flags some of them ($\tau = 85$ detects 14% in the top decile), but only by accepting namesake alerts on everyone who shares those names. The trade-off is not an artefact of the calibration: name-only evidence about a common name is weak evidence, and the only sound repair is a secondary identifier.

5.5 Matcher families

Composition amplification is a property of the wrapper, not of the scorer, and decoy competition is a wrapper too. Table 3 repeats the analysis for six matchers at operating points that raise about the same alert volume. Every one of them concentrates alerts on the same populations: the ratio between the largest and the median country false-alert rate is 7.5 for the token-sorted indel ratio, 7.0 for Jaro-Winkler, 9.8 for the token-set ratio, 5.0 for a phonetic matcher, 11.0 for a character n-gram TF-IDF cosine and 6.7 for a multilingual sentence embedding, with Gini coefficients between 0.47 and 0.61. A better scorer buys accuracy, not equality. Replacing the threshold by a lowered threshold plus decoy competition at the same volume lowers the disparity for the exhaustively screened matchers (7.5 to 6.0 and 7.0 to 4.8) and raises recall (66.3% to 68.6% and 56.2% to 61.9%).

The neural and LLM matchers are run in the deployed two-stage form, where retrieval keeps at most twenty candidates per list and the decoy universe is only twice the size of the list, which truncates the competition in (4); their improvement is therefore a lower bound, as the two string matchers show when they are run both ways (rows marked † against their exhaustive counterparts).

6 Related work

One-to-many identification. The mechanism is established in face biometrics, where a demographic's false-positive identification rate grows with its presence in the gallery, so equalising pairwise false match rates equalises burden only for a balanced gallery (National Institute of Standards and Technology, 2024; Howard et al., 2019; Sirotin & Vemury, 2021; Grother et al., 2019). We transport that law to name screening, where the per-customer driver is a namesake count, measure it on a real sanctions list, and supply the calibration the biometric literature leaves open.

Table 3: Matcher families at matched alert volume. *max/median* is the ratio of the largest to the median country false-alert rate. Recall under the decoy rule is given with the threshold's recall in brackets. † marks the two-stage setting, where retrieval truncates the decoy competition.

	threshold			threshold + TDC, same volume		
matcher	volume (%)	max/median	Gini	recall (%)	max/median	Gini
token-sorted indel	6.3	7.5	0.561	68.6 (66.3)	6.0	0.489
Jaro-Winkler	5.1	7.0	0.548	61.9 (56.2)	4.8	0.440
token-sorted indel†	6.1	10.2	0.598	64.7 (63.5)	10.2	0.595
Jaro-Winkler†	6.1	8.8	0.580	58.6 (56.3)	6.8	0.514
token-set†	6.5	9.8	0.607	61.4 (67.8)	7.7	0.576
phonetic (Metaphone)†	6.5	5.0	0.467	60.0 (58.0)	5.2	0.481
char n-gram TF-IDF†	6.2	11.0	0.604	67.6 (67.8)	7.5	0.571
multilingual embedding†	4.9	6.7	0.566	36.3 (40.2)	5.6	0.561

Matching, fairness and calibration. Frequency-aware record linkage has long known that common names carry less evidence (Fellegi & Sunter, 1969; Winkler, 1989; 2014; Christen, 2012), and recent work audits entity matching for fairness across groups and calibrates scores to make thresholds comparable (Shahbazi et al., 2023; Karakasidis & Pitoura, 2019; Moslemi & Milani, 2024; Liu & Frasincar, 2025); frequency adjustment is known to shrink but not remove ethnoracial gaps in linkage quality (Lam et al., 2026). All of this operates on pairs and equalises over groups that must be named and observed; our guarantee is per customer, hence for every group at once, uses no attribute, and concerns the one-to-many maximum rather than a pairwise score. Our oracle per-country threshold is the strongest member of that family, since it equalises burden exactly, and Table 2 shows what it costs in detection. Belin & Rubin (1995) calibrate false-match rates by mixture modelling, a parametric route to the same question.

Competition-based error control. Target-decoy competition originates in proteomics (Elias & Gygi, 2007; 2010) and has a substantial theory (He et al., 2015; Keich et al., 2015; Levitsky et al., 2017; Emery et al., 2020; Luo et al., 2023), closely related to knockoffs (Barber & Candès, 2015; Candès et al., 2018; Etourneau et al., 2023) and to rank-based Monte-Carlo testing (Besag & Clifford, 1989; Phipson & Smyth, 2010; Westfall & Young, 1993). Robach et al. (2025) bring it to record linkage, injecting synthetic records to estimate one global false-discovery proportion for a linkage output; we instead run a competition per customer inside a one-to-many screening decision, with composition-matched decoy lists, and bound each individual's and hence each group's alert probability. The p-value is a conformal-style outlier test (Vovk et al., 2005; Bates et al., 2023) and the robustness analysis follows the conformal treatment of non-exchangeability (Barber et al., 2023); the guarantee is of the kind group-conditional conformal methods (Vovk, 2012; Romano et al., 2020) obtain only for prespecified groups, in the spirit of fairness without demographics (Hashimoto et al., 2018; Lahoti et al., 2020; Hébert-Johnson et al., 2018).

Sanctions screening. Recent work benchmarks matchers for screening, including LLMs, on labelled *pairs*: Allen & Hatfield (2025) compare LLMs with fuzzy matching on sanctions name pairs, and Smith et al. (2026) release a large pair-labelled entity-matching benchmark built from the same data source we use. Pairwise accuracy does not determine who bears the alerts in deployment; our benchmark measures the one-to-many regime where the list's composition, not the matcher's pairwise quality, sets the burden.

Limitations and outlook. Our customers are non-sanctioned politically exposed persons: public, verifiable and country-labelled, but not a bank's book, and officials are not a random sample of customers. The stratification is the country recorded for a listed person, a crude proxy for a name population, and 22% of listed persons carry no country at all. Our decoy generators are deliberately simple; the benchmark exists so that better ones, including character-level and LLM-based name models, can be scored against the exchangeable reference. We screen names only, while production systems also hold dates of birth and identifiers, and Proposition 5 says that is where the detection of common-name listed persons has to come from. The guarantee is a statement about a sampling model for the list, in the sense in which a permutation test is a statement about one realised sample. Two

conclusions survive all of this. Alert burden in one-to-many screening is a property of the list and of the customer's name population, so a better matcher cannot fix it and institutions should measure it; and when burden is calibrated rather than thresholded, the price is paid on common names, where a name-only decision should not have been made in the first place.

REPRODUCIBILITY STATEMENT

Every number in the paper is produced by the released pipeline from two dated public snapshots, and every figure and table is generated from files written by that pipeline (Appendix B lists dataset versions, checksums, seeds and runtimes). The proofs of all five propositions are in Appendix A. The code, the decoy generators, the frozen query and universe files, and the scripts that regenerate every table and figure are in the supplementary material; the full benchmark runs on two CPU cores in under three hours.

ETHICS STATEMENT

This study uses two public, redistributable datasets about public roles: consolidated sanctions lists and a register of politically exposed persons. It processes names of real people, which is unavoidable when studying name screening, and we take three precautions. We never infer ethnicity, religion or any protected attribute from a name: the only grouping variable is the country that the source data already records, and it is used solely to measure who bears alerts, never as an input to a screening rule. The procedure we propose is demographic-blind by construction, which is also why it is compatible with legal regimes that forbid differential treatment by national origin. We do not release any new personal data: the benchmark ships entity identifiers, code and dated snapshot references, so that the underlying records stay under the provider's own licence and are retrieved from the provider.

We also considered the dual-use question. The identifiability limit of Proposition 5 makes explicit that a listed person with a common name is hard to detect by name alone. This is not new information to sanctions evaders, and it is well known to compliance practitioners; stating it precisely is what justifies the use of secondary identifiers instead of lower thresholds, and lower thresholds are what produce the disparate burden we document. On balance we judge the transparency to be beneficial. Finally, false alerts are not a cost only to banks: people who are repeatedly flagged face delayed payments, account closures and de-risking, and the burden falls hardest on customers whose names are common in populations that sanctions lists over-represent. That asymmetry is the motivation for this work.

AI USE STATEMENT

In this work we used generative AI tools (Anthropic's Claude, operated as a coding and research agent by the author) for research ideation and experiment design, for implementing the methods and running all experiments, for retrieval and discovery of related work, for formulating the mathematical claims and drafting the proofs, and for drafting and polishing the text of this paper. We did not use generative AI tools to generate the data: the datasets are public snapshots, and the synthetic decoy names are produced by the released statistical generators, which are ordinary code and use no language model.

Every part of this was verified before it entered the paper. The bibliography was checked entry by entry against Crossref, DataCite, PubMed or the publisher's own page, and entries that could not be verified were dropped rather than cited; no reference in this paper was accepted on an LLM's assertion. The propositions were checked line by line against the definitions and the proofs rewritten where the argument was loose; where an early draft of a claim was wrong (an initial form of Proposition 1(ii) compared group means rather than distributions) it was corrected to the statement proved in Appendix A. Every number in the text is generated mechanically from experiment output files rather than transcribed, and the pipeline was re-run end to end to confirm it. The author reviewed the code, the experimental design, the outputs and the claims, and takes responsibility for the final content of this work, including text, claims and artefacts produced with the aid of generative AI.

REFERENCES

Jeffrey S. Allen and Max S. S. Hatfield. Can LLMs improve sanctions screening in the financial system? Evidence from a fuzzy matching assessment. Finance and Economics Discussion Series 2025-092, Board of Governors of the Federal Reserve System, 2025.

Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015. doi: 10.1214/15-AOS1337.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023. doi: 10.1214/22-AOS2244.

Thomas R. Belin and Donald B. Rubin. A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90(430):694–707, 1995. doi: 10.1080/01621459.1995.10476563.

Julian Besag and Peter Clifford. Generalized Monte Carlo significance tests. *Biometrika*, 76(4):633–642, 1989. doi: 10.1093/biomet/76.4.633.

Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: 'Model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B*, 80(3):551–577, 2018. doi: 10.1111/rssb.12265.

Peter Christen. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Data-Centric Systems and Applications. Springer, 2012. doi: 10.1007/978-3-642-31164-2.

Joshua E. Elias and Steven P. Gygi. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 4(3):207–214, 2007. doi: 10.1038/nmeth1019.

Joshua E. Elias and Steven P. Gygi. Target-decoy search strategy for mass spectrometry-based proteomics. In *Proteome Bioinformatics*, volume 604 of *Methods in Molecular Biology*, pp. 55–71. Humana Press, 2010. doi: 10.1007/978-1-60761-444-9_5.

Kristen Emery, Syamand Hasam, William Stafford Noble, and Uri Keich. Multiple competition-based FDR control and its application to peptide detection. In *Research in Computational Molecular Biology (RECOMB 2020)*, volume 12074 of *Lecture Notes in Computer Science*, pp. 54–71. Springer, 2020. doi: 10.1007/978-3-030-45257-5_4.

Lucas Etourneau, Nelle Varoquaux, and Thomas Burger. Unveiling the links between peptide identification and differential analysis FDR controls by means of a practical introduction to knockoff filters. In *Statistical Analysis of Proteomic Data*, volume 2426 of *Methods in Molecular Biology*, pp. 1–24. Springer, 2023. doi: 10.1007/978-1-0716-1967-4_1.

Ivan P. Fellegi and Alan B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969. doi: 10.1080/01621459.1969.10501049.

Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test (FRVT) part 3: Demographic effects. Technical Report NISTIR 8280, National Institute of Standards and Technology, 2019.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1929–1938, 2018.

Kun He, Yan Fu, Wen-Feng Zeng, Lan Luo, Hao Chi, Chao Liu, Lai-Yun Qing, Rui-Xiang Sun, and Si-Min He. A theoretical foundation of the target-decoy search strategy for false discovery rate control in proteomics. *arXiv preprint arXiv:1501.00537*, 2015. URL https://arxiv.org/abs/1501.00537.

Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1939–1948, 2018.

John J. Howard, Yevgeniy B. Sirotin, and Arun R. Vemury. The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pp. 1–8. IEEE, 2019. doi: 10.1109/BTAS46853.2019.9186002.

Alexandros Karakasidis and Evaggelia Pitoura. Identifying bias in name matching tasks. In *Proceedings of the 22nd International Conference on Extending Database Technology (EDBT)*, 2019. doi: 10.5441/002/edbt.2019.72.

Uri Keich, Attila Kertesz-Farkas, and William Stafford Noble. Improved false discovery rate estimation procedure for shotgun proteomics. *Journal of Proteome Research*, 14(8):3148–3161, 2015. doi: 10.1021/acs.jproteome.5b00081.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 728–740, 2020.

Joseph Lam, Mario Cortina-Borja, Ruth Blackburn, and Katie Harron. Mitigating ethnoracial-based linkage bias with feature and frequency-aware methods. *International Journal of Population Data Science*, 11(5), 2026. doi: 10.23889/ijpds.v11i5.3618.

Lev I. Levitsky, Mark V. Ivanov, Anna A. Lobas, and Mikhail V. Gorshkov. Unbiased false discovery rate estimation for shotgun proteomics based on the target-decoy approach. *Journal of Proteome Research*, 16(2):393–397, 2017. doi: 10.1021/acs.jproteome.6b00144.

Yuan Liu and Flavius Frasincar. FairNM: Fairness in name matching. In *Natural Language Processing and Information Systems (NLDB 2025)*, volume 15837 of *Lecture Notes in Computer Science*, pp. 36–46. Springer, 2025. doi: 10.1007/978-3-031-97144-0_4.

Dong Luo, Arya Ebadi, Kristen Emery, Yilun He, William Stafford Noble, and Uri Keich. Competition-based control of the false discovery proportion. *Biometrics*, 79(4):3472–3484, 2023. doi: 10.1111/biom.13830.

Mohammad Hossein Moslemi and Mostafa Milani. Threshold-independent fair matching through score calibration. In *Proceedings of the Conference on Governance, Understanding and Integration of Data for Effective and Responsible AI (GUIDE-AI)*, pp. 40–44. ACM, 2024. doi: 10.1145/3665601.3669845.

National Institute of Standards and Technology. Relating 1:1 results to 1:N applications. FRTE demographics report annex, 2024. URL https://pages.nist.gov/frvt/reports/demographics/implications_for_1N.pdf.

Belinda Phipson and Gordon K. Smyth. Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 39, 2010. doi: 10.2202/1544-6115.1585.

Kayané Robach, Michel H. Hof, and Mark A. van de Wiel. False discovery estimation in record linkage. *Statistics in Medicine*, 44(23–24), 2025. doi: 10.1002/sim.70292.

Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel J. Candès. With malice toward none: Assessing uncertainty via equalized coverage. *Harvard Data Science Review*, 2020. doi: 10.1162/99608f92.03f00592.

Nima Shahbazi, Nikola Danevski, Fatemeh Nargesian, Abolfazl Asudeh, and Divesh Srivastava. Through the fairness lens: Experimental analysis and evaluation of entity matching. *Proceedings of the VLDB Endowment*, 16(11):3279–3292, 2023. doi: 10.14778/3611479.3611525.

Yevgeniy B. Sirotin and Arun R. Vemury. Demographic variation in the performance of biometric systems: Insights gained from large-scale scenario testing. EAB Demographics in Biometric Systems Workshop, DHS S&T Maryland Test Facility, 2021. URL `https://mdtf.org/publications/EAB2021-Demographics.pdf`.

Chandler Smith, Magnus Sesodia, Friedrich Lindenberg, and Christian Schroeder de Witt. OpenSanctions Pairs: Large-scale entity matching with LLMs, 2026. arXiv:2603.11051.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pp. 475–490, 2012.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. doi: 10.1007/b106715.

Peter H. Westfall and S. Stanley Young. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley, 1993. ISBN 0471557617.

William E. Winkler. Frequency-based matching in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods*, pp. 778–783. American Statistical Association, 1989.

William E. Winkler. Matching and record linkage. *WIREs Computational Statistics*, 6(5):313–325, 2014. doi: 10.1002/wics.1317.

A Proofs

A.1 Proposition 1

Under $\mathcal{M}(n, \pi, F)$ the entries $y_1, \dots, y_n$ are i.i.d. from F_π, so with $\mu = \mu_\pi(x, \tau) = \sum_g \pi_g q_g(x, \tau) = \Lambda/n$,

$$\mathbb{P}[A_\tau = 0] = \prod_{i=1}^{n} \mathbb{P}[s(x, y_i) < \tau] = (1 - \mu)^n,$$

which gives (3); the lower bound uses $(1 - \Lambda/n)^n \le e^{-\Lambda}$ and the upper bound is the union bound $\mathbb{P}[A_\tau = 1] \le \sum_i \mathbb{P}[s(x, y_i) \ge \tau] = \Lambda$. Also $N_\tau(x, L) \sim \mathrm{Bin}(n, \mu)$, so $\mathbb{E}N_\tau = \Lambda$, and $\Lambda \mapsto 1 - (1 - \Lambda/n)^n$ is strictly increasing on $[0, n]$.

(i) If $q_h(x, \tau) = 0$ for $h \neq g(x)$ then $\Lambda(x, \tau) = n\pi_{g(x)} q_{g(x)}(x, \tau)$. Taking expectations over $x \sim P_g$ and using $1 - e^{-\Lambda} \le \mathbb{P}[A_\tau = 1] \le \Lambda$,

$$\mathrm{FAR}_g(\tau) = n\pi_g \bar{q}_g(\tau)\big(1 + O(n\pi_g \bar{q}_g(\tau))\big),$$

so if $\bar{q}_a(\tau) = \bar{q}_b(\tau)$ the ratio $\mathrm{FAR}_a/\mathrm{FAR}_b$ tends to π_a/π_b as the rates vanish.

(ii) $\mathbb{P}[A_\tau = 1 \mid x] = h(\Lambda(x, \tau))$ with h increasing, so stochastic dominance of $\Lambda(\cdot, \tau)$ under P_a over P_b implies $\mathbb{E}_{P_a} h \ge \mathbb{E}_{P_b} h$ for every τ, i.e. $\mathrm{FAR}_a(\tau) \ge \mathrm{FAR}_b(\tau)$ on T; hence no $\tau \in T$ equalises them. □

A.2 Proposition 2

Fix x and write $S_0, \dots, S_K$ for the screening scores against $L, D_1, \dots, D_K$. Each S_j is the same fixed measurable function of the corresponding list, so Assumption 1 makes $(S_0, \dots, S_K)$ exchangeable. Define the conservative ranks $R_j = 1 + |\{i \neq j : S_i \ge S_j\}|$, so that $p(x) = R_0/(K + 1)$.

Claim: for every realisation and every $r \in \{1, \dots, K + 1\}$, $|\{j : R_j \le r\}| \le r$. Order the distinct values of $\{S_j\}$ decreasingly as $v_1 > v_2 > \dots$, and let c_k be the number of indices j with $S_j \ge v_k$.

Every j in the tie group at v_k has $R_j = c_k$ exactly. Hence $\{j : R_j \le r\}$ is the union of the tie groups whose cumulative count c_k is at most r, and the size of that union is the largest such c_k, which is at most r. This proves the claim.

Taking expectations and using exchangeability, $\mathbb{E}\,\mathbf{1}\{R_0 \le r\} = \frac{1}{K+1}\sum_j \mathbb{E}\,\mathbf{1}\{R_j \le r\} \le \frac{r}{K+1}$. Since p takes values in $\{1/(K+1), \dots, 1\}$, for any $\alpha \in [0,1]$ we have $\mathbb{P}[p \le \alpha] = \mathbb{P}[R_0 \le \lfloor \alpha(K+1) \rfloor] \le \lfloor \alpha(K+1) \rfloor / (K+1) \le \alpha$. All statements are conditional on x, so they hold on every event of the form $\{x \in G\}$; requiring additionally $S_0 \ge \tau_{\min}$ intersects the alert event with another event and can only lower its probability. □

A.3 Proposition 3

Conditionally on x, write $p = \Psi(S_0; S_1, \dots, S_K)$ with Ψ non-increasing in its first argument, and let $f(t) = \mathbb{P}[\Psi(t; S_1, \dots, S_K) \le \alpha]$, which is well defined because $S_1, \dots, S_K$ are independent of S_0, and non-decreasing in t. Then $\mathbb{P}_P[p \le \alpha] = \int f \, d\mathcal{L}_P(S_0)$ and likewise for Q. Since $0 \le f \le 1$,

$$\left| \int f \, d\mathcal{L}_P - \int f \, d\mathcal{L}_Q \right| \le d_{\mathrm{TV}}(\mathcal{L}_P, \mathcal{L}_Q).$$

When $S_0 \sim Q$ the $K+1$ scores are i.i.d. and Proposition 2 gives $\int f \, d\mathcal{L}_Q \le \alpha$, which yields (6). Monotonicity of f gives the stochastic-dominance statement, and Corollary 4 follows by applying it group by group at a fixed overall budget. □

A.4 Proposition 5

Conditionally on x and on $S_0 = \sigma$, the indicators $\mathbf{1}\{S_j \ge \sigma\}$, $j \in [K]$, are i.i.d. Bernoulli$(\rho(\sigma))$, so $|\{j : S_j \ge S_0\}| \sim \mathrm{Bin}(K, \rho(\sigma))$ and $p \le \alpha \iff \mathrm{Bin}(K, \rho(\sigma)) \le \alpha(K+1) - 1$. The law of large numbers gives the limit $\mathbf{1}\{\rho(\sigma) < \alpha\}$, and $\rho(\sigma) = 1 - (1 - \Lambda_D/n_D)^{n_D} \approx 1 - e^{-\Lambda_D(x,\sigma)}$ by Proposition 1 applied to a decoy list. □

B Data, seeds and compute

Snapshots. OpenSanctions *sanctions* collection, version `20260919014701-llq` (2026-09-19 01:47:01 UTC), file `targets.simple.csv`, SHA-256 `b4b8c463689e7b17235ed6d2b253066bfcb3f467accaa108dcd06728ceea5493`; OpenSanctions *peps* collection, version `20260918152701-kum` (2026-09-18 15:27:01 UTC), file `targets.simple.csv`, SHA-256 `2c98b7a54423180556d0091d05be3ea9e1628e49981c288f04a3641366c425e5`. Both are distributed by the provider under CC BY-NC 4.0; we redistribute code and identifiers only.

Normalisation. `unidecode` romanisation, lower-casing, removal of non-letters, removal of single-character tokens, and a requirement of at least two remaining tokens. The screening key is the token-sorted name, for which the indel ratio equals `token_sort_ratio` on the raw name (verified exactly on a sample).

Splits and seeds. Customer sample: at most $1{,}500$ persons per country and never more than half of a country's non-listed persons, seed `20260920`; the remainder is the decoy pool. Detection queries: $8{,}000$ listed persons with at least two spellings, one spelling held out per person, same seed. Decoy list realisations use seeds recorded in the released scripts; $K = 199$ throughout unless stated.

Compute. Two CPU cores, 8 GB RAM, no GPU. The exhaustive screening passes evaluate 1.4×10^8 name pairs per second with an early-exit cutoff; the full pipeline (four universes, five matchers, all experiments) runs in under three hours.

C Identity leakage check

A customer who is in fact the listed person would be counted as a false alert. Of 40,755 customers, 302 reach an exact normalised-name match with the list. Among those with a birth year recorded on both sides, the birth years agree for 5.8% against a 1.48% agreement rate for random pairs of listed

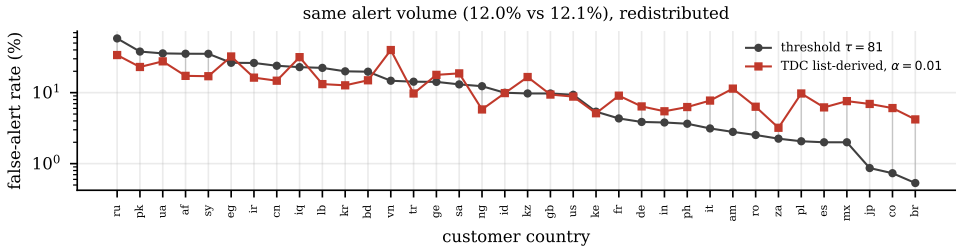

Figure 3: Country false-alert rates under a threshold and under target-decoy competition at the same alert volume. Calibration moves burden from the countries the list over-represents to countries whose names it does not.

Table 4: Detection against namesake density. Λ_D is the expected number of names in one composition-matched decoy list that reach the query's own match score. TDC at $\alpha = 0.01$ with list-derived decoys.

decile of Λ_D	1	2	3	4	5	6	7	8	9	10
median Λ_D	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.60
TDC recall (%)	100	100	100	100	100	100	100	89	0	0
$\tau = 85$ recall (%)	83	82	82	81	81	82	82	76	32	14

and non-listed birth years, so the excess consistent with genuine identity is on the order of a handful of customers, i.e. below 0.02% of the population and below 2% of the exact-match alerts.

D Additional results

E Decoy generators

All generators are ordinary statistical code; no language model is involved.

Component resampling. For a stratum c we fit, on the names of that stratum in the donor source, the distribution of token counts and the token distributions at the first, middle and last positions. A generated name draws a token count and then one token per position, rejecting repeats inside a name. Strata with fewer than 100 donor strings fall back to a global fit; this affects 1.6% of generated strings.

Alias structure. Sanctions lists record several spellings per person. We match this in two steps. First, every decoy person draws a target number of name strings from the empirical distribution of name counts per listed person in its stratum. Second, the extra strings are generated by a four-component mixture: a near-duplicate (one character-level transliteration edit), a moderate variant (two to four edits), a partial variant (one token of the base name kept, the rest resampled), and a fresh name from the stratum. The mixture weights are fitted by non-negative least squares so that the similarity histogram between generated variants and their base name matches the list's own alias-to-primary similarity histogram, measured in four bands (< 50, 50–70, 70–90, ≥ 90). On the list the target histogram is $(0.14, 0.29, 0.28, 0.29)$ and the fitted generator reproduces $(0.17, 0.26, 0.28, 0.28)$ for pep donors and $(0.20, 0.24, 0.24, 0.32)$ for list-derived donors.

Realisation of a list. Every list in the experiments, including the screened list in the validation of Section 5.2, is realised by the same mechanism: each sampling unit of the universe (a name string, or a person with all their strings) is included independently with a stratum-specific probability chosen so that the expected stratum size matches the target. Because the screened list and the decoy lists are realised by the same mechanism in the validation, exchangeability there is exact rather than approximate; on the real list the screened list is fixed and only the decoys are realised this way, which is one of the discrepancies Proposition 3 bounds.

Sources. PEP: non-listed politically exposed persons, sampled per stratum to match the list's person count. PEP-UNIF: the same pool with a single global inclusion rate. PEP+VAR: pep donors with the

alias structure above, capped at ten donors per listed person and stratum. LIST-DERIVED: decoy persons generated entirely from the list's own per-stratum token distributions with the same alias structure, ten decoy persons per listed person; generated names that collide with a real listed name are dropped (0.9% of generated strings).

F MATCHERS AND THE TWO-STAGE PROTOCOL

The full-scale matchers are exhaustive: every query is scored against every string of every universe, with no blocking. TSR is the indel ratio on the token-sorted name (equal to `token_sort_ratio`); JW is Jaro-Winkler on the token-sorted name; TSET is the token-set ratio; PHON replaces every token by its Metaphone code before the token-sorted indel ratio. Neural and LLM matchers use the deployed two-stage design: stage one retrieves with TSR at a permissive cutoff of 55 and keeps the 20 best candidates per query and per universe; stage two rescores those candidates. The candidate file is frozen and released, so any matcher can be scored on identical retrieval. The embedding matcher is `paraphrase-multilingual-MiniLM-L12-v2` with cosine similarity mapped to $[0, 100]$. The LLM matcher scores each candidate pair with the instruction: *"You compare personal names for a sanctions screening system. For each numbered pair, state the probability (0-100) that the two names could refer to the same person, allowing for transliteration, spelling variants, name order and dropped middle names."* at temperature 0, in batches of 25 pairs. Two-stage retrieval means a neural or LLM matcher cannot alert on a pair that stage one did not retrieve; this is how such systems are deployed, and it applies identically to the real list and to the decoy lists, so the competition in (4) remains between like and like.

G LABEL-FREE MOMENT MATCHING OF DECOYS

The operator can observe, for every stratum of the list, how many customers have a best match of at least the operating threshold among the entries of that stratum, and a correct decoy model must reproduce those counts. Solving for per-stratum inclusion probabilities by bisection on that criterion uses only the list's own strata and no customer attribute. In the validation setting of Section 5.2 it improves the aggregate and leaves the tail: the overall alert rate falls from 0.85α to 0.61α and the share of countries above the nominal level from 33% to 22%, while the worst country moves only from 3.8α to 3.7α. The reason is visible in the score tail: real name populations repeat whole names far more often than any independent-component generator does.

H ROBUSTNESS

Number of decoy lists. In the validation setting of Section 5.2 with pool-synthetic decoys, the worst country's rate is 2.4α at $K = 49$, 2.9α at $K = 199$ and 2.9α at $K = 999$: the calibration error is a property of the decoys, not of the number of them, and $K = 199$ is enough for $\alpha = 0.05$.

Identity leakage. Dropping every customer whose best list match scores at least 95 and whose recorded birth year agrees with that listed person (10 customers) leaves the headline unchanged: the alert rate at $\tau = 85$ moves from 6.33% to 6.31% and the largest-to-median country ratio from 7.5 to 7.5.

Sampling uncertainty. Bootstrapping the customer sample (400 resamples) gives 95% intervals at $\tau = 85$ of [36.5, 41.3]% for the worst country's rate, [6.6, 8.8] for the largest-to-median ratio and [0.55, 0.58] for the Gini coefficient, so the disparity is far larger than sampling noise.

Mixing decoy sources. An operator could take all available real non-listed donors for a stratum and top the stratum up with synthetic names, which combines the donors' near-duplicate tail with unlimited supply. It does not close the gap: the worst country's rate is 3.2α against 3.21α for the best single source, because the residual error is in which customers have namesakes, not in how many names the decoy list holds.