When "Anonymous" Data Stops Being Anonymous: Measuring How Public Data Accumulates into Re-Identification Risk

AUTHOR

Tirth Joshi, M.S. in Artificial Intelligence

FACULTY ADVISOR

Aaron Ross, MBA

ACKNOWLEDGEMENTS

I thank Aaron Ross for mentorship and guidance, and the Katz School of Science and Health for supporting this work.

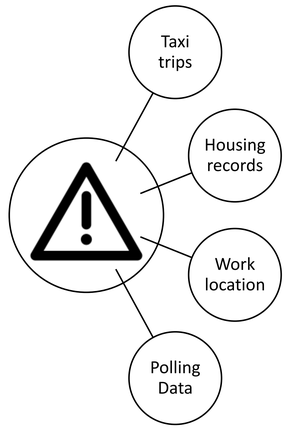
SUBMISSION STATUS

This paper is currently under review for presentation at the Annual Privacy Forum (APF 2026), Austria.

Katz School of Science and Health

The Problem: Anonymization Is Not Enough

Privacy does not fail only because of one bad dataset; it fails when many “safe” datasets can be linked together.



The Puzzle-Piece Effect

No single "anonymized" dataset may identify anyone. But when multiple datasets are combined, the picture sharpens and individuals can be uniquely singled out.

Better Question

Does this release make people easier to identify when combined with other available data?

Why Existing Metrics Fall Short

No existing metric measures cumulative identifiability pressure across releases without requiring access to sensitive microdata or live re-identification attacks.

Solution: Re-Identification Pressure Index (RPI)

RPI is an auditable, ecosystem-level metric of structural identifiability pressure. It answers: how close is this information environment to a state in which uniqueness and therefore identification becomes routine?

RPI: The metric we introduce

Population scale: The same profile sparsity is more dangerous in a small population than a large one. RPI adjusts for this.

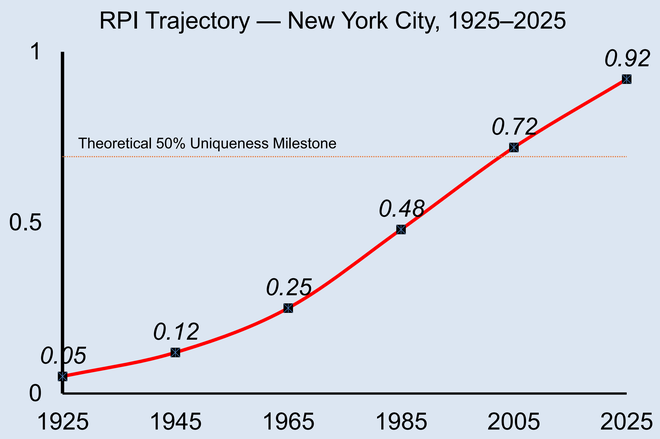
$$RPI = H_2(X) - \log N = -\log(N \cdot C)$$

Rényi-2 entropy: A precise measure of how many effectively distinguishable profiles exist. Higher entropy = sparser profiles = higher pressure.

Collision probability: The chance that two individuals share the same observable profile. High collisions = more privacy. Low collisions = more exposure.

RPI increases when public profiles become more unique relative to population size.

Results



Identifier	6 → 26
Entropy	9.2 → 20.5
Public record Coverage	35% → 95%

- Under regular occupancy conditions, when RPI → +∞, virtually all individuals are unique. The benchmark RPI ≈ ln 2 ≈ 0.693 marks the ~50% uniqueness transition.
- Six anchor years were analyzed using only publicly available data, no privileged access, no private databases. The question: how close to routine identifiability was the public information environment in each era?

Central Finding:
RPI rises **monotonically** across all six periods and remains directionally robust under all tested modeling assumptions including the most conservative.

Takeaways

1. A better question

Instead of asking only “Is this dataset anonymous?”, this work asks: How much does a new release, integration, or procurement decision increase re-identification pressure in the broader data environment?

2. Why it matters

A dataset may appear safe on its own but still increase risk when combined with other public or administrative data. This means privacy review should evaluate not only the dataset itself, but also its effect on the whole information ecosystem.

3. Practical use

RPI can support decisions about:
- data release
- data sharing
- integration across sources
- procurement and vendor review
- privacy-by-design assessment

Preprint & References