

Workshop on data security and privacy

Banco de Portugal and Irving Fisher Committee · Session 5 · Lisbon, 15 September 2026

Quantitative risk assessment for data privacy

Probability models, estimation uncertainty and governance thresholds

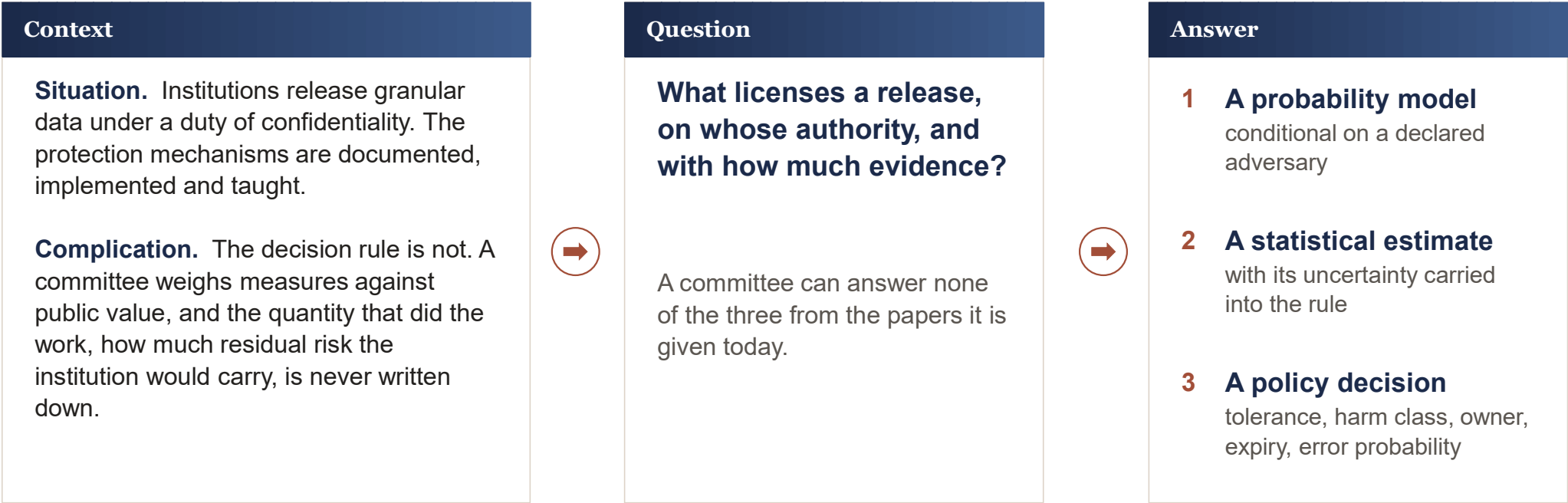
Ronald Rethabile Malamule Mahomane

Department of Computing, Mathematical and Statistical Sciences, University of Namibia, Windhoek

School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg

Basis Energy, Victoria, Australia

A release decision has three parts, and most of the difficulty in practice comes from collapsing them



Two failures follow from collapsing them: a committee that treats a point estimate as the risk releases at its own boundary about half the time, and an institution that treats a protection parameter as a tolerance never states the probability it accepted.

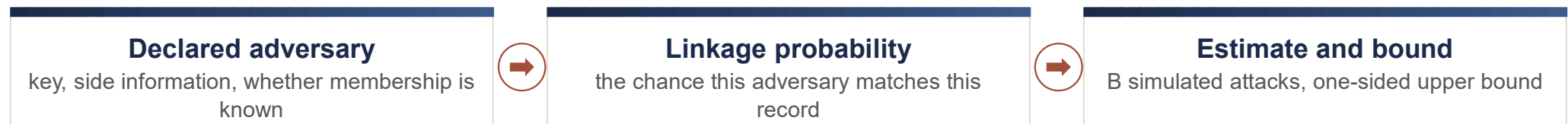
Three stages: what the adversary faces, what the institution computes, and what it then decides

THE WHOLE ARGUMENT ON ONE PAGE

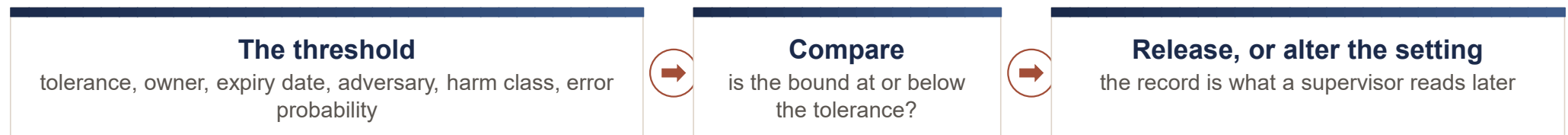
1 THE RELEASE



2 THE MEASUREMENT



3 THE DECISION



Two things the three stages do not settle. A programme of releases needs its own joint simulation, because the marginal figures cannot bound what an adversary achieves by pooling them. And a formal privacy mechanism runs a separate accountant, with its own budget and its own owner, because its parameter is not a linkage probability.

The familiar measures are theorems about one linkage experiment, and k-anonymity is a decision rule in disguise

A The measures are theorems

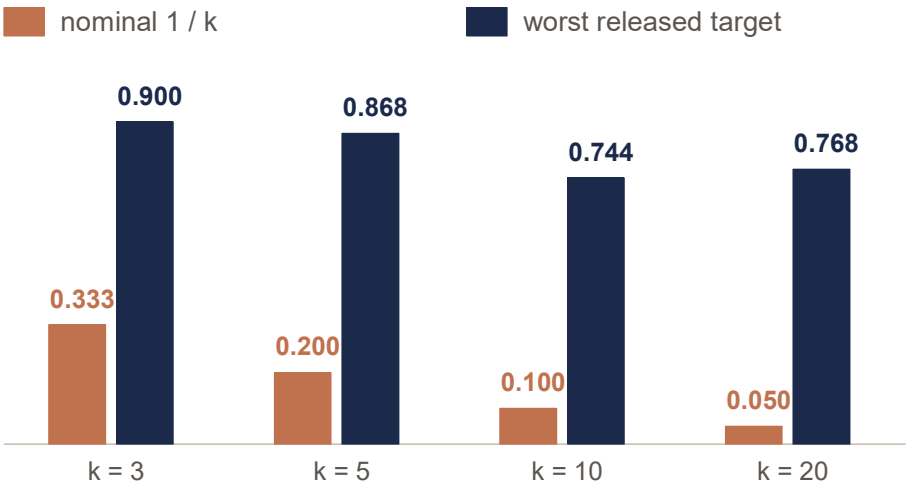
PART A · SECTIONS 3 AND 3.4 · PROPOSITIONS 1 AND 2

FOUR ASSUMPTIONS BEHIND THE RECIPROCAL FREQUENCY

Unperturbed key		HOLDS
Key observed without error	with reporting error	FAILS
Records in a class exchangeable	with a correlated covariate	FAILS
Uniform selection among them		HOLDS

Two of the four fail whenever the adversary holds informative side information, and the posterior linkage probability then exceeds $1 / f k$. On the census sample 86.3 per cent of records are unique on their key, and the declared adversary reaches 0.3540 where exact matching reports 0.1093.

k-ANONYMITY IS THE RULE $\sup R \leq 1/k$ UNDER THE SAME FOUR



A 10-anonymous release holds a record identified with probability 0.74, bounded below by 0.64, and even the average over released targets, 0.1129, exceeds the nominal 0.100.

First negative result. A k-anonymous release does not deliver a risk of $1/k$, and the same reading applies to l-diversity and t-closeness.

The figure is an estimate, so the tolerance and the standard of evidence are one decision

B The figure is an estimate

PART B · SECTIONS 4 AND 8 · PROPOSITIONS 3, 4 AND 13

ONE FRONTIER SETTING AGAINST A TOLERANCE OF 0.12

point estimate 0.1058, B = 1,200



A rule that releases on the point estimate releases at its own boundary about half the time, so a tolerance of 0.12 should reject this release. And no observed success is not evidence of a small risk: with B successful attacks the exact one-sided bound is about 3/B.

And the two frequency measures are not interchangeable

Sample-frequency risk dominates population-frequency risk pathwise, so the direction of the error is known before any data are seen: 0.466 against 0.064 in the simulation, a factor of 7.3.

WHAT IT COSTS TO CERTIFY A TOLERANCE

Tolerance	Fixed sample	If you may stop	Ratio
0.01	299	984	3.3
0.001	2,995	12,417	4.1
0.00001	299,572	not feasible	—

Watching your own simulation is not free

An institution that runs 4,000 attacks, looks at its bound twenty times and stops when it first falls below the truth does so wrongly in 20 per cent of 500 runs at a nominal 5 per cent. The anytime-valid sequence does it in none, and costs 1.06 times the width. An adversary that gains ten times what a false allegation costs it acts at a posterior of 0.136, not one half.

An institution adopting a very small tolerance has committed to a large simulation programme, or to a model in place of simulation.

The tolerance is a policy quantity with an owner and an expiry date, and it is implied rather than chosen

C The tolerance is policy

$$G = (\tau, o, d, a, h, \alpha)$$

τ tolerance	the residual risk the institution will carry
o named owner	without it the tolerance belongs to nobody
d expiry date	without it a 2019 assessment governs a 2026 release
a adversary	without it the number has no referent
h harm class	a tolerance for nuisance applied where the harm is physical
α error probability	how much evidence the institution requires

(5) AND THE TOLERANCE IS NOT FREE

Price of harm κ	Implied tolerance	Records kept
0.25	0.464	1.00
0.50	0.345	1.00
1.00	0.145	0.94
5.00	0.081	0.58

At the optimum the admissible risk is the point where the frontier slope equals the price the institution puts on harm relative to public value. Doubling that price from 0.5 to 1 cuts the tolerance from 0.345 to 0.145. A uniform tolerance across releases is efficient only where the frontier slope is common to them, which it generally is not.

The target set has to be explicit too: a maximum over records, a maximum over a protected subgroup and an average over a population are three different controls.

Marginal risks cannot bound what an adversary who pools the releases achieves

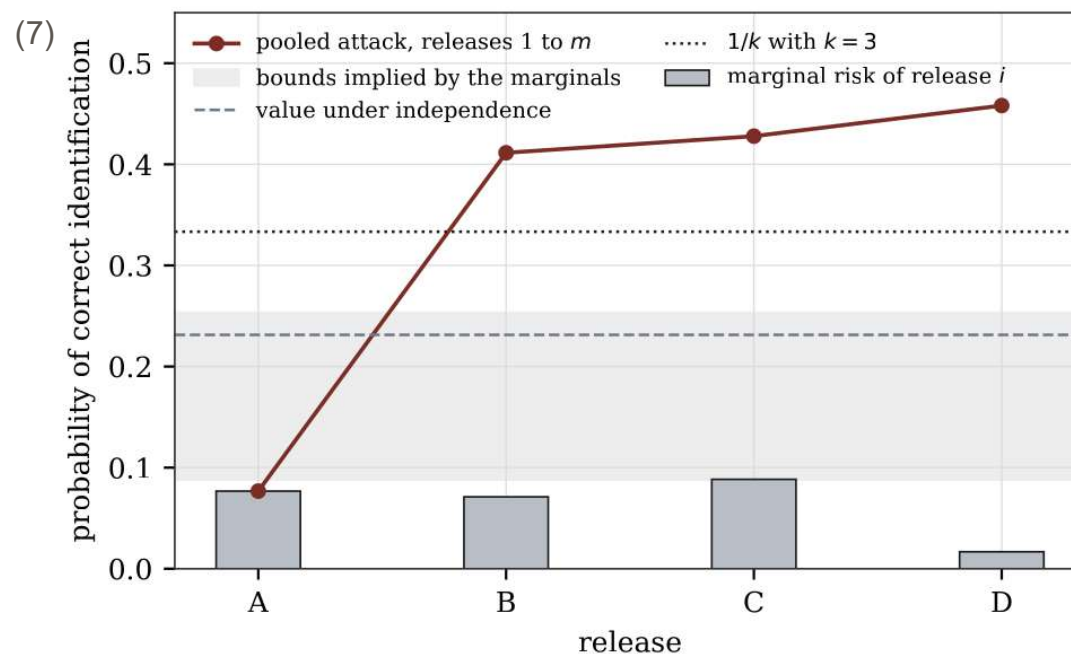
PART C · SECTION 7 · PROPOSITIONS 9 AND 10

$$\max_i R_i \leq R_U \leq \min(1, \sum_i R_i)$$

Both ends of the marginal bound are attained, so the marginals identify no narrower interval. But the union is the wrong event: for every tolerance and every number of releases there exist releases with marginal risk below it whose pooled attack succeeds with probability one.

Four ordinary releases from one sample

Each 3-anonymous, with no unique record anywhere. Marginals 0.0768, 0.0711, 0.0884 and 0.0168, so the bound is [0.0884, 0.2530]. Two releases already reach 0.4114 and four reach 0.4581, which is 1.8 times that bound. On a real census file the same comparison gives 1.86 times.



Ranking releases by their own risk does not rank their contribution: release D has the smallest marginal and adds the most to the pool.

Second negative result. Marginal figures license no programme claim, and a programme tolerance needs direct joint simulation.

Every claim reappears on files that were actually published, and six tests could have contradicted it



EVIDENCE · SECTIONS 11 AND 14

86.3%

of 29,501 census records are unique on their full key

0.98

identifies the worst record in a five-anonymous release, against a nominal 0.2

1.86

times the union bound is what four pooled releases from that file reach

SIX TESTS, AND WHAT CAME BACK

Linkage calibration	a gap of 0.007 correct, 0.053 misspecified	PASSES
Population-frequency validation	the fitted law is within 1 per cent	PASSES
Adversary stress test	a sharper covariate lifts risk by 40 per cent	INFORMS
Mechanism monotonicity	every row breaches the nominal bound	FAILS
Composition	the marginal bound is beaten at two releases	FAILS
Temporal transport	cannot be run at assessment time	PENDING

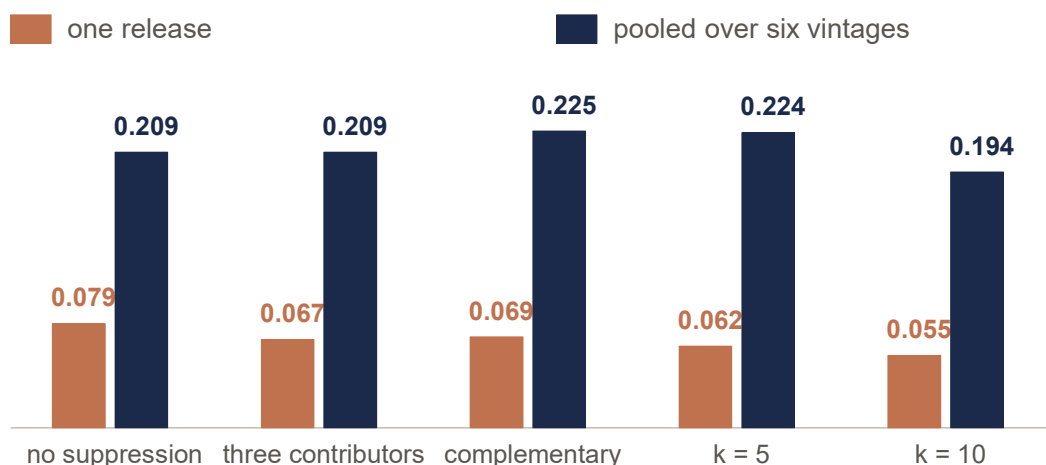
What the evidence does not settle

The magnitudes on any particular register, because those depend on the auxiliary data an adversary actually holds. The evaluation is a lower bound on what is possible: a better adversary exists, and every policy faces the same one.

Reproducing the equations in software is an implementation check. Evidence for the scenario model has to be capable of contradiction, and two of the six tests fail.

Seven policies agencies operate, put against four attacks on one monthly credit register

THE CASE · SECTIONS 12 AND 13



A credit register creates five harms at once: a borrower identified, a relationship alleged, arrears status inferred, a contribution reconstructed from totals, and two vintages differenced. One minimum cell count governs the first and moves the others at four different rates.

- 1 Every policy reduces it, none removes it**
the binding constraint is the exposure reading, not the cell size
- 2 The programme is a different story**
0.062 at $k = 5$ becomes 0.224 across the six vintages
- 3 The adversary is accurate when it acts**
it acts on 0.011 of attempts and is right on 0.761 of those
- 4 The average conceals a subgroup**
the worst size class carries a corrected bound of 0.061

Same tolerance of 0.12: calibrated on one release the answer is a cell count of one, keeping 32,581 relationships. Calibrated on the pooled attack it is fifty, keeping 20,389.

The record to file with every release decision, each stage written down before the next

RECOMMENDATION · SECTION 15

- 1 Declare** the attack scenario, the target set, the harm class, the uncertainty set and the loss
- 2 Compute** the point estimate, the one-sided bound and the number of replicates behind it
- 3 Review** whether the declared scenario is plausibly conservative, using the six tests
- 4 Apply** the tolerance, through the named owner, to the bound rather than to the point estimate
- 5 File** the record with its expiry date and evidence sufficient to reproduce the computation

For a programme, two further items	Two separate responsibilities
The joint attack simulation with its own tolerance, and the accountant's running total where a formal mechanism is used.	The owner owns the residual risk. The analyst owns the correctness of the calculation. They should sit with separate people.

Donald Rethabile Mahomane, ron@besicongroup.com — the full owner carries the proofs and the pipeline

The record does not remove judgement. It puts judgement where it belongs and leaves the calculation reproducible.