

Lossy Distance-Shell Tomography on Paths: Recovery from Mean and Covariance

Lennart Rudolph 

October 11, 2026

Abstract

Distance-shell sensors report how many targets lie at each hop, without identifying them or separating branches. We show that independent relay failures can make these aggregates more informative. Under nearest-surviving-ancestor bypass, for every known survival probability $0 < p < 1$, the mean and covariance of an interior path sensor's report determine the two integer-valued leg populations up to exchange. Means alone retain exactly the noiseless shell information on every rooted tree. Recovery needs only the covariance diagonal and one adaptively selected column. For legs of length n and counts at most h , a variance-sensitive rounding estimator has an explicit exponential bound at fixed p . At the designed probability $p = 2n/(2n + 1)$, two independent batches give $O(h^4 n \log(1/\delta))$ rounds with $O(n)$ scalar accumulators. A binary two-point lower bound proves the matching minimax order $\Theta_h(n \log(1/\delta))$ for each fixed h and $0 < \delta \leq 1/4$.

For nonempty populations with $0 < p < 1$, the full law of the total reported count determines an unknown survival probability and the multiset of positive vertex counts. For nonempty binary populations, the mean and covariance already calibrate the survival probability; an explicit counterexample shows why this need not hold for general integer counts. On a spider with 2^k legs, parity populations agree in all joint moments through order k but differ at order $k + 1$. Endpoint reports connect the model to deletion-channel trace reconstruction, giving imported super-polynomial lower and quasi-polynomial upper bounds at fixed p . Joint landing formulas and a recursive generating function describe how shared ancestry enters the law on general trees. We conjecture that this law identifies the labelled occupied core, with exact finite evidence for every $0 < p < 1$. Selected endpoint and two-leg results, together with variance, deterministic rounding and testing components of the sample analysis, are formalized in Lean 4; the full probability guarantee remains a written proof.

1 Introduction

Distance-shell tomography asks what labelled sensors can recover about an anonymous integer population from its distance histograms [1, 2]. A sensor knows how many targets lie at each distance, but not which vertices carry them. Distinct populations with equal reports differ by an invisible integer trade. For a sensor inside a path, the ambiguity is especially simple: at each depth it sees the sum of the populations on its two sides.

We ask what happens when the relays used to measure hop distance fail. In each independent round, every relay survives with known probability p . Failed relays are bypassed, surviving descendants attach to their nearest surviving ancestor, and targets on a failed relay are silent. At the tree level, this is the tree-edit-distance deletion operation studied by Davies, Rácz and Rashtchian [16, Section 1]. Here the sensor retains only the resulting hop histogram, discarding

the surviving tree and its branch identities. Hop-based localization and tree-based data collection motivate this aggregate observable [18, 19]; the cited protocols do not assert that physical rerouting implements this exact contraction rule.

The main result is that failures resolve the two-sided ambiguity. For every $0 < p < 1$, the mean and covariance determine the two leg populations up to exchange. This does not follow from a better mean signal: on every rooted tree, the mean contains exactly the noiseless shell information. The additional information lies in correlations between targets whose reports depend on shared relays.

Why repeated aggregate distances? A count field specifies where load or occupancy is concentrated without requiring individual target identities. Tree-based aggregation of counts and histograms is a familiar sensor-network primitive [25]. The earlier distance-shell papers explain why locating individual vertices does not suffice to reconstruct such a population, and develop the corresponding integer-trade and sensor-placement theory [1, 2]. Here the sensor remains fixed while repeated relay masks change which populations contribute and at which hops. The question is whether the resulting fluctuations reveal spatial information that a single noiseless histogram suppresses, and how many rounds are needed to read it. An anonymous report is an observation constraint, not a formal privacy guarantee.

Failures act on the hidden spatial population *before* aggregation. In general, the lossy law cannot be generated from the noiseless histogram alone by a population-independent noise channel. Equal noiseless histograms can produce different lossy laws without contradicting the data-processing principle. Shared relay ancestry creates the correlations that distinguish them. Using fluctuations of aggregate measurements to identify hidden quantities is an established idea in network tomography [26, 27]. The endpoint and interior cases here distinguish two effects: failures can create identifiability where it was absent, while making an already identifiable problem statistically harder.

Results.

- (i) *Two-leg recovery* (Theorem 5.4 and Propositions 5.5 and 5.6). From the mean and covariance we recover the unordered pair of leg words for every $0 < p < 1$. The reconstruction is explicit. For integers $n, h \geq 1$ and $0 < \delta < 1$, it recovers legs of length n with counts at most h , with probability at least $1 - \delta$, from

$$2000h^4p^{-2n}(1-p)^{-1}\log(4n^2/\delta)$$

rounds, rounded up to an integer. Thus $\exp(O(n))$ rounds suffice for fixed p , fixed confidence and bounded counts. Only n diagonal covariance entries and at most $n - 1$ entries of one adaptively chosen column are needed for exact recovery (Corollary 5.9). If p can be chosen, setting $p_n = 2n/(2n + 1)$ and using two independent batches yields the sharper bound

$$n \log(1/(2\delta)) \leq N^*(n, h, \delta; p_n) \leq 2 \lceil 16000h^4n \log(1 + 8/\delta) \rceil,$$

for $n \geq 2$, $h \geq 1$ and $0 < \delta \leq 1/4$. Here N^* is the minimax fixed round budget. The upper bound uses $O(n)$ scalar accumulators (Theorem 5.14 and Corollary 5.15).

- (ii) *Means on trees* (Theorem 4.1). The mean report is an invertible binomial transform of the noiseless histogram. Any two populations with the same noiseless report remain indistinguishable by first moments.

- (iii) *Calibration without geometry* (Proposition 4.3 and Corollary 4.4). For a nonempty population, the law of the total reported count determines p and, for $p < 1$, the multiset of positive vertex counts. For binary populations, first and second moments already determine p , so the two-leg recovery theorem extends to unknown survival probability. This extension fails for unrestricted integer counts without further information (Example 4.5).
- (iv) *Endpoint complexity* (Theorems 3.2 and 3.3). A report together with the number of surviving relays is exactly a deletion-channel trace. The cited trace-reconstruction results [5–7] imply an $m^{\Omega(\log \log m)}$ round lower bound for binary populations supported on the first m relays at each fixed $0 < p < 1$. Quasi-polynomially many rounds in the full path length suffice at fixed p : plain reports suffice for binary populations at known rational p , and horizon-aware reports suffice for arbitrary counts.
- (v) *An exact moment obstruction* (Theorem 6.1). On a spider with 2^k legs, parity populations agree in all joint moments through order k , but, for $0 < p < 1$, their first-hop moments of order $k + 1$ differ by the explicit nonzero quantity in (11). No fixed moment order identifies populations on all spiders, even within pairs whose full laws differ.
- (vi) *A tree conjecture* (Conjecture 6.2). For $0 < p < 1$, we conjecture that the full law determines the occupied core as a rooted integer-labelled tree. Exact computation verifies this at every $p \in (0, 1)$ for binary populations on trees with at most eight non-root vertices, and for counts at most 2 on trees with at most five. The two-leg theorem proves the path case.

The covariance mechanism. Write $c = a + b$ for the fold and $d = a - b$ for the difference of the leg populations. The mean determines c . If $Q(u)$ is the covariance array of one leg, independence gives $\text{Cov}(R) = Q(a) + Q(b)$, and its quadratic dependence on u gives

$$Q(d) = 2 \text{Cov}(R) - Q(c).$$

For $0 < p < 1$, the deepest nonzero diagonal entry of $Q(d)$ identifies the last nonzero coordinate of d and its magnitude. After choosing its sign, one covariance column determines every earlier coordinate by a triangular system with pivots $1 - p^k$. All those pivots are nonzero for $0 < p < 1$ and vanish at $p = 1$. This is how failures create the information missing from the fold.

Scope and prior work. The population model, shell reports and invisible-trade viewpoint are inherited from [1, 2]. The first paper develops the integer observation kernel and optimal grid sensing. The second studies erasure or corruption of whole sensor reports, deterministic moment compression and equalization. Its erasure model preserves all distances in retained reports. Its distributed interpretation assumes reliable internal aggregation, and its final open question 6 asks about partial loss within aggregation [2, Section 11 and open problems]. We analyze one specified form of internal loss: a failed relay’s population is silenced, while surviving descendants use contracted paths.

The two uses of “moment” also concern different observables. In [2, Theorem 21], a sensor s compresses a fixed population to the distance power sum $M_j(s) = \sum_v x(v) d(s, v)^j$. Here we use $\mathbb{E} \prod_{t=1}^j R_A(l_t)$, a joint moment of random hop-bin counts across relay masks at one sensor. Both settings admit the classical Boolean parity trade [20, 21]; the spider result applies its cancellation to independent leg cumulants and computes the first nonzero moment gap. The parity pattern itself is inherited.

The contribution developed here is the exact two-leg covariance recovery theorem for these superimposed reports, together with its adaptive sparse interface and an explicit finite-sample bound. The proof combines a channel-specific triangular covariance argument with standard quadratic algebra; the sample analysis combines hop-dependent variance bounds with Bernstein’s inequality [4] and a two-point testing lower bound. Total-count calibration and shared-ancestor formulas supply supporting consequences. The elementary binomial calibration identity already appears in deletion-channel inference [30, Lemma 4.1], and the spider obstruction uses the classical parity construction cited above. These results and their proofs are independent of the imported trace-reconstruction bounds. The endpoint complexity statements in Section 3.1 and their interior lower-bound consequence (Corollary 5.8) use [5–7]; the binary plain-report upper bound also uses the cited decoder’s specific observables. The general-tree statement remains a conjecture.

What changes in version 2. This is a small v2 update of the original document dated October 10, 2026 [3], focused on the sampling analysis. The previous version used a Hoeffding bound [22] of order $h^4 p^{-2n} (1-p)^{-2} \log(n/\delta)$ and obtained a quadratic designed round budget. The new analysis retains the factor $1-p^k$ in the variance at hop k , uses matching hop-dependent detection thresholds, and sums the hop-specific tail probabilities after independent sample splitting. This gives the matching order at $p_n = 2n/(2n+1)$ in Corollary 5.15, resolving the quadratic-sharpness subquestion of the earlier Problem 9.2 and supplying a linear-storage guarantee for its Problem 9.6. The fixed-probability complexity, optimal dependence on h , and optimization over survival schedules remain open.

Statistical network tomography. Vardi [27] estimates source–destination traffic intensities from repeated aggregate link counts under Poisson traffic and deterministic or Markovian routing. First and second moment equations supply additional constraints beyond mean counts. Duffield and Lo Presti [26] use end-to-end delay covariance to infer shared-path delay variance and logical tree structure. Chen, Cao and Bu [28] study distributional identifiability from linear mixtures of independent latent variables, and Lev-Ari, Ephraim and Mark [29] extend traffic-rate inference to higher cumulants. These works establish the broader use of fluctuations and higher moments to resolve aggregate inverse problems. Our unknowns are fixed integer vertex populations, and an independent relay mask changes their hop-bin assignments in each round. The recovery theorem below resolves this particular inverse problem on paths.

Deletion traces and population recovery. Trace reconstruction was introduced by Batu, Kannan, Khanna and McGregor [8]. Worst-case upper bounds improved from $\exp(\tilde{O}(\sqrt{n}))$ traces [9] to $\exp(O(n^{1/3}))$ [10, 11], which is optimal for mean-based methods [10], and then to $\exp(\tilde{O}(n^{1/5}))$ [12]. Burudgunte, Valiant and Wang obtained the first quasi-polynomial bound [13]; the quasi-polynomial bounds we use are the refinements of [6, 7]. Lower bounds were polynomial, $\tilde{\Omega}(n^{5/4})$ [14] and $\tilde{\Omega}(n^{3/2})$ [15], until the super-polynomial bound of [5]. Davies, Racz and Rashtchian [16] reconstruct the labels of a known tree from traces that reveal the entire surviving labelled, ordered tree. For binary populations, the report studied here is a depth-histogram projection of their tree-edit-distance trace; natural-number populations extend the labels to multiplicities. The projection discards branch order, individual labels and the number of surviving unoccupied vertices at each depth.

Ban, Chen, Freilich, Servedio and Sinha [17] study population recovery under the deletion channel. Narayanan [31] develops higher-moment trace estimators and uses Prony’s method to recover information about a sparse distribution over strings. In both works, each observation

first selects one latent string and then applies deletions. A spider hub here instead observes the coordinatewise sum of independent, zero-padded traces of all its legs in the same round. This superposition leads to different moment equations from those of mixture recovery.

Hartung, Holden and Peres [30] study known position-dependent deletion probabilities under regularity assumptions, as well as possibly unknown symbol-dependent probabilities. Their mean-trace transform and binomial calibration argument are direct precedents for the corresponding elementary ingredients below. The heterogeneous-relay and calibration questions in Section 9 concern what can be recovered after branchwise traces have been superimposed.

Quadratic recovery and moment obstructions. Each entry of the one-leg covariance $Q(d)$ is a quadratic form in the leg difference $d = a - b$, hence a linear function of the entries of $dd^\top$. Encoding a vector modulo sign by its outer product of rank at most one is the familiar lifting viewpoint used in phase retrieval; see Candès, Eldar, Strohmer and Voroninski [32, Section 2.2]. The proof below establishes injectivity directly from the triangular structure of the deletion covariance. For several legs, the standard additivity of joint cumulants [33, Section 2.7.1] combines with classical parity cancellation [20, 21]. The stochastic spider theorem identifies the resulting first unequal moment explicitly.

Formalization. Eleven selected statements are proved in Lean 4 with Mathlib: the endpoint-trace bridge, noiseless and first-moment identities for paths, two-leg recovery from moments or covariance, the exact law-equivalence classification, and an explicit broken trade. Additional library theorems certify recovery from a covariance diagonal and one column, invariance of the law-equivalence classes across survival probabilities, and the zero-survival boundary. The formalization uses finite sums over failure patterns and an arbitrary linearly ordered field. Version 2 adds real-valued finite-sample components: actual-report variance bounds, conditional deterministic recovery, randomized-test lower bounds, and the geometric error budget. Bernstein concentration and the complete probability guarantee are not formalized. The precise scope and exclusions are given in Section 7.

2 Lossy hop-count reports

Throughout, $p \in (0, 1]$ is the survival probability and $q = 1 - p$. Unless explicitly stated otherwise, reconstruction assumes that p and the ambient rooted tree (in particular the path lengths) are known. Proposition 4.3 and Corollary 4.4 treat unknown p . The complete-invariant conjecture also compares laws on different ambient trees.

Definition 2.1 (Lossy report). Let T be a finite tree rooted at the sensor ρ , put $V^\circ = V(T) \setminus \{\rho\}$, write $d(v)$ for the depth of v , and write $u \preceq v$ when $u \in V^\circ$ lies on the path from ρ to v . A *population* is a function $x: V^\circ \rightarrow \mathbb{N}$. A *round* draws an alive set $A \subseteq V^\circ$ containing each vertex independently with probability p . For $v \in A$ let $\delta_A(v) = \#\{u \in A : u \preceq v\}$. The *report* of the round is

$$R_A(k) = \sum_{v \in A, \delta_A(v)=k} x(v) \quad (k \geq 1),$$

and $\mu_{T,x,p}$ denotes its law.

Contracting every failed vertex into its parent turns T into a tree on $\{\rho\} \cup A$ in which v has depth $\delta_A(v)$. This is the tree-edit-distance deletion operation of [16, Section 1], followed here by

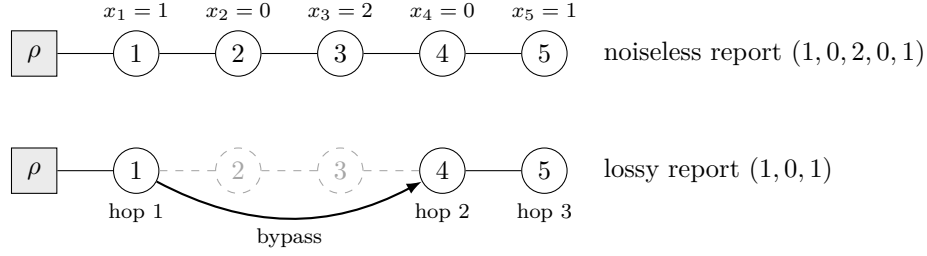


Figure 1: An endpoint sensor on a path. Relays 2 and 3 fail, so the two targets on relay 3 fall silent, and relays 4 and 5 are heard at hops 2 and 3. The alive relays carry the deletion trace $(1, 0, 1)$ of the word $(1, 0, 2, 0, 1)$.

aggregation of surviving labels by depth. We assume this contraction is the effective routing rule. At $p = 1$ the report is the noiseless histogram $N_x(k) = \sum_{d(v)=k} x(v)$, the distance-shell report of a sensor at ρ in the sense of [1]. On a network that is not a tree we take, as a modelling assumption, that hop counts are collected along a fixed routing tree and that every relay below a failed relay is re-attached to its nearest surviving ancestor. Real rerouting around a failure may instead find a different path, and that is not modelled here.

Lemma 2.2 (Landing law). *A vertex v of depth d is alive and heard at hop distance k with probability*

$$P_d(k) = \binom{d-1}{k-1} p^k q^{d-k} \quad (1 \leq k \leq d), \quad P_d(k) = 0 \quad (k > d).$$

Proof. The vertex is alive with probability p , and independently $\delta_A(v) - 1$ counts the alive vertices among its $d - 1$ proper ancestors in V° , a $\text{Bin}(d - 1, p)$ variable. $\square$

Lemma 2.3 (Branches are independent). *If $T_1, \dots, T_r$ are the subtrees hanging from the children of ρ , then $R_A = \sum_j R^{(j)}$, where $R^{(j)}$ is the contribution of T_j , measured from ρ , and the summands are independent. Hence $\mu_{T,x,p}$ is the convolution of the branch laws.*

Proof. For $v \in T_j$, $\delta_A(v)$ depends only on $A \cap T_j$, and the sets $A \cap T_j$ are independent. $\square$

The *occupied core* of (T, x) consists of the root together with the union of the root paths to vertices with $x(v) > 0$, labelled by x on its non-root vertices. The empty population has the root-only core. Vertices outside the core carry no targets and are not ancestors of occupied vertices, so R_A depends only on $A \cap \text{core}$.

Proposition 2.4. *If the occupied cores of (T, x) and (T', x') are isomorphic as rooted $\mathbb{N}$ -labelled trees, then $\mu_{T,x,p} = \mu_{T',x',p}$ for every p .*

Proof. Couple the alive indicators on the two occupied cores through the rooted label-preserving isomorphism. Their joint distribution is unchanged because the indicators are independent with the same survival probability. The isomorphism preserves each occupied vertex's surviving-ancestor count and label, hence its contribution to the report. Vertices outside the cores neither contribute targets nor lie above an occupied vertex, so their indicators do not affect either report. $\square$

For paths we use words. A sensor at the *end* of a path sees one leg carrying $x = (x_1, \dots, x_n)$, where x_i targets sit on the relay at true distance i . A sensor in the *interior* sees two legs carrying $a, b \in \mathbb{N}^n$; a shorter leg is padded with unoccupied relays, which changes nothing because relays deeper than every target never affect the plain report. Exchanging the padded leg words is the ambiguity considered below. For unequal ambient leg lengths, this need not be an automorphism of the full known path; it is reflection of the occupied rooted path. On a path with an endpoint sensor the *horizon* is $|A|$, the current hop length of the leg, and the *horizon-aware report* is the pair $(R_A, |A|)$. Padding a leg with empty relays preserves its plain report law, but can change its horizon-aware law.

2.1 How shared ancestry enters second moments

The one-vertex landing law depends only on depth. Two-vertex landing probabilities also depend on where the routes meet. For vertices u, v , define the bivariate polynomial

$$B_{uv}(z, w) = \sum_{k, l \geq 1} \mathbb{P}\{u, v \in A, \delta_A(u) = k, \delta_A(v) = l\} z^k w^l.$$

Writing $[z^k w^l]B$ for a coefficient, these polynomials give every second moment:

$$\mathbb{E}R_A(k)R_A(l) = \sum_{u, v \in V^\circ} x(u)x(v)[z^k w^l]B_{uv}(z, w). \quad (1)$$

Proposition 2.5 (Joint landing kernels). *Let $a = d(u)$ and $b = d(v)$. For $u = v$,*

$$B_{uu}(z, w) = pzw(q + pzw)^{a-1}.$$

If u is a strict ancestor of v , then

$$B_{uv}(z, w) = p^2zw^2(q + pzw)^{a-1}(q + pw)^{b-a-1}.$$

The reverse ancestor case follows by exchanging u, v and z, w . If neither vertex is an ancestor of the other and their lowest common ancestor has depth c , then

$$B_{uv}(z, w) = p^2zw(q + pzw)^c(q + pz)^{a-c-1}(q + pw)^{b-c-1}.$$

The root has depth 0 and is not a random relay.

Proof. An ancestor shared by both routes contributes a factor $q + pzw$, while a proper ancestor belonging to only one route contributes $q + pz$ or $q + pw$. In the diagonal case, the survival of u contributes pzw . If u is a strict ancestor of v , their simultaneous survival contributes p^2zw^2 : the survival of u increments both hop counts, and that of v increments only its own. There are $a - 1$ common proper ancestors and $b - a - 1$ intervening relays. For incomparable vertices, their simultaneous survival contributes p^2zw , with c shared non-root ancestors and the stated numbers of private ancestors. Multiply the factors using independence. Expanding the product of the two report sums proves (1). $\square$

Thus covariance records weighted pair information about depth and shared ancestry. When u, v belong to different root branches, $c = 0$ and the joint kernel factors into the two marginal kernels, as required by branch independence. These formulas make explicit which geometric information first moments discard; they do not assert that second moments determine every tree population.

3 Endpoint reports and deletion traces

An endpoint sensor has no branch ambiguity: each surviving relay occupies a distinct hop. Relay failure therefore acts by deleting entries of the population word. The only information a plain report omits is the number of surviving empty relays after its last nonzero entry. This distinction controls which trace-reconstruction results transfer directly.

Proposition 3.1 (Endpoint reports are traces). *Let $A = \{i_1 < \dots < i_t\}$ be the alive relays of a path with an endpoint sensor and population word $x \in \mathbb{N}^n$, and let $\tau(A) = (x_{i_1}, \dots, x_{i_t})$ be the deletion trace. Then $R_A(k) = \tau_k$ for $k \leq t$ and $R_A(k) = 0$ for $k > t$, so*

$$\tau(A) = (R_A(1), \dots, R_A(|A|)).$$

The horizon-aware report and the complete trace determine each other, and the plain report is the trace with its trailing zeros removed. Consequently the horizon-aware report of x has the law of a complete deletion trace of the word x with retention probability p .

Proof. $\delta_A(i_r) = r$, so the targets of relay i_r are counted in entry r . Every relay survives independently with probability p , which is the deletion channel. $\square$

3.1 Round complexity

Fix the survival probability and report type. For a class $\mathcal{X}$ of populations and $s \in (0, 1]$, let $N_s(\mathcal{X})$ be the least number of independent rounds from which some estimator recovers every $x \in \mathcal{X}$ with probability at least s , with unrestricted computation; set this number to ∞ if no finite budget works. Following [5], $T_{q,s}(m)$ denotes the least number of complete traces with which some estimator recovers every $y \in \{0, 1\}^m$ with probability at least s .

Theorem 3.2 (Failures cost super-polynomially many rounds). *Let $n \geq m$, and let $\mathcal{X}_m$ be the 0/1 populations on the path P_n supported on the m relays nearest the endpoint sensor; they have mass at most m . For plain and for horizon-aware reports,*

$$N_s(\mathcal{X}_m) \geq T_{q,s}(m).$$

Consequently, by [5, Theorem 1.1], which holds for every success probability $s \in (0, 1]$, for each fixed $q \in (0, 1)$ and $s \in (0, 1]$ recovery needs $m^{\Omega(\log \log m)}$ rounds. More precisely, for every $0 < c < 1/(4 \log 2)$ and along every sequence (m, q) with $q^3 \log m \rightarrow \infty$, it needs at least $m^{c \log(q^3 \log m)}$ rounds eventually.

Proof. Let an estimator recover $\mathcal{X}_m$ from N horizon-aware reports with probability at least s . Given N independent complete traces $\tau^1, \dots, \tau^N$ of an unknown $y \in \{0, 1\}^m$, form $R^t(k) = \tau_k^t$ for $k \leq |\tau^t|$ and $R^t(k) = 0$ otherwise, and $H^t = |\tau^t| + B^t$ with independent $B^t \sim \text{Bin}(n - m, p)$. For the population $x = (y, 0^{n-m})$, the alive relays among the first m produce a trace of y , and the remaining relays only enlarge the horizon. By Proposition 3.1, (R^t, H^t) therefore has exactly the law of a horizon-aware report of x . Feeding these reports to the estimator recovers x , hence y , with probability at least s , so $N \geq T_{q,s}(m)$. Plain reports are functions of horizon-aware reports, so the bound transfers. The asymptotic statements are [5, Theorem 1.1]. $\square$

Moreover, for each fixed $q \in (0, 1)$, [5, Theorem 1.2] gives a sequence of pairs of distinct 0/1 populations on P_n whose single-round horizon-aware report laws have total-variation distance $n^{-\omega(1)}$ as $n \rightarrow \infty$.

Theorem 3.3 (Quasi-polynomially many rounds suffice). *For an endpoint sensor on P_n with known $n \geq 2$ and known survival probability p :*

- (a) (Efficient recovery [6, Theorem 1.1].) *There are absolute constants $C, c \geq 1$ and a randomized algorithm that, for every rational $p \in (0, 1]$ and every 0/1 population x , requests at most $M = \lceil \exp(C B(n, p)) \rceil$ plain reports and returns x with probability at least $2/3$, within $(n + L + M)^c$ bit operations, where L is the bit length of p ,*

$$B(n, p) = p^{-1}(\log n) X^2(1 + \log X)^6, \quad X = 1 + \frac{\log n}{1 + \log(1/q)} \quad (q > 0),$$

and $X = 1$ if $q = 0$. For fixed p both resources are quasi-polynomial; if $q \leq n^{-\kappa}$ for a fixed $\kappa > 0$ the sample count is polynomial in n and the running time is polynomial in $n + L$. The same guarantees hold with horizon-aware reports.

- (b) (Arbitrary counts [7, Theorem 1].) *For every $p \in (0, 1]$, there is an estimator, depending on n and p , that recovers every population $x \in \mathbb{N}^n$ with probability at least $2/3$ from $\exp(O(p^{-1}(\log n)^3(1 + \log \log 2n)^6))$ horizon-aware reports, with an absolute constant. This is a sample-complexity statement.*

For fixed p , in both parts the number of rounds is quasi-polynomial in the path length n , not in the size m of the support: populations supported on $m \ll n$ relays are covered with n in place of m .

Proof. For part (a), [6, Theorem 1.1] is stated for complete binary traces, but its decoder uses only the observables described in its Sections 2.1 and 5.1: for nonempty rank sets $I \subseteq \{1, \dots, n\}$ it estimates $\prod_{k \in I} \tau_k$, with value zero if a rank is absent. By Proposition 3.1, this is exactly $\prod_{k \in I} R_A(k)$, computable from the plain report. Thus the empirical inputs to that decoder, and hence its success and complexity guarantees, are unchanged. At $p = 1$, pad the plain report to the known length n and read x . This plain-report consequence uses the cited algorithm, beyond the report-trace identity alone.

For part (b), N horizon-aware reports are N independent complete deletion traces of the word x over the alphabet of its values. Apply [7, Theorem 1], which explicitly includes general symbols observed exactly. Its general-symbol construction first collects a set of observed symbols, then applies binary tests to candidate words over that set. The estimator is thus chosen from n and p before the unknown population; it does not require its alphabet as prior information. $\square$

Remark 3.4. Without failures the endpoint sensor reads x in a single round, and by Theorem 4.1 the mean lossy report still determines x . Failures therefore do not destroy identifiability; they make it expensive, and by Theorem 3.2 no polynomial number of rounds suffices. Plain reports, which hide how many relays survive beyond the last heard target, still reveal every mean-based statistic: the plain report padded with zeros is the trace padded with zeros, so $\mathbb{E}R_A(k)$ is the padded mean trace. The mean-based algorithms of [10, 11] therefore recover every 0/1 population from $\exp(O(n^{1/3}))$ plain reports at each fixed q . Theorem 3.3(a) improves this to quasi-polynomially many plain reports at fixed rational p , using higher-order products. A horizon is needed to identify a report with the complete trace, but not to evaluate these particular trace statistics. The optimal rate remains open (Problem 9.1).

4 What the mean retains

A target's marginal landing probability depends only on its true depth, not on how its branch is attached to the rest of the tree. Averaging therefore erases the branch dependence created by failures.

The resulting triangular transform is the familiar mean-trace transform from deletion-channel reconstruction [10, 11]; see also the position-dependent formula in [30, Lemma 2.1]. Grouping its coefficients by original depth gives the following tree statement.

Theorem 4.1 (Means). *Let T be a rooted tree, x a population and N_x its noiseless report. For every $p \in (0, 1]$ and $k \geq 1$,*

$$\mathbb{E} R_A(k) = \sum_{d \geq k} N_x(d) \binom{d-1}{k-1} p^k q^{d-k}.$$

The map $N_x \mapsto \mathbb{E} R_A$ is injective. Hence two populations, on possibly different trees, have equal mean lossy reports if and only if they have equal noiseless reports.

Proof. By linearity and Lemma 2.2, $\mathbb{E} R_A(k) = \sum_v x(v) P_{d(v)}(k)$; grouping vertices by depth gives the formula. The matrix $(P_d(k))_{k,d}$ vanishes for $d < k$ and has diagonal entries $P_k(k) = p^k \neq 0$, so the noiseless report is recovered from the bottom entry upward. $\square$

Corollary 4.2. *Every trade invisible to a noiseless sensor remains invisible to the mean of its lossy report. In particular, statistics of first order alone cannot exploit failures.*

For an interior path sensor, the noiseless report is $N = a + b$. The mean therefore recovers the folded population exactly, but cannot distinguish two different splittings of the same fold.

4.1 Total counts, calibration and the positive label multiset

Summing over hops removes the geometric effect of bypass. Put $Y = \sum_{k \geq 1} R_A(k)$ and let $B_v = [v \in A]$. Then

$$Y = \sum_v x(v) B_v, \quad \mathbb{E} Y = p \sum_v x(v), \quad \text{Var}(Y) = pq \sum_v x(v)^2. \quad (2)$$

Every surviving vertex contributes its count exactly once, regardless of its surviving ancestors. These identities separate a geometry-free calibration problem from the spatial inverse problem.

Proposition 4.3 (Calibration and positive counts from the total law). *Let $x \not\equiv 0$ be a population on any finite rooted tree, with an unknown $p \in (0, 1]$. The law of Y determines p and the total population $M = \sum_v x(v)$ through*

$$M = \max \text{supp}(Y), \quad p = \frac{\mathbb{E} Y}{M}.$$

If $0 < p < 1$, it also determines the multiset of positive values of x , including multiplicities. Consequently, for nonempty two-leg populations with $0 < p < 1$, the full report law determines both p and the unordered pair of padded legs.

Proof. List the positive counts as $w_1, \dots, w_r$. The corresponding survival indicators are independent, and

$$F(t) := \mathbb{E} t^Y = \prod_{i=1}^r (q + p t^{w_i}).$$

Its degree is $M = \sum_i w_i$: all occupied relays survive with probability $p^r > 0$. Its derivative at 1 is $\mathbb{E} Y = pM$, proving the calibration formula even when $p = 1$.

For $p < 1$, set $\lambda = p/q > 0$ and normalize

$$H(t) = F(t)/F(0) = \prod_i (1 + \lambda t^{w_i}).$$

If $H \neq 1$, its least positive exponent m is the smallest remaining count. Only one-factor choices contribute to t^m , so its coefficient is λ times the multiplicity of m . Recover that integer, divide by the corresponding power of $1 + \lambda t^m$, and repeat. Each division strictly lowers the degree, and the process ends at 1. This reconstructs the positive counts uniquely. With p now known, apply Theorem 5.4 to the full two-leg law. $\square$

The empty population gives $Y = 0$ for every p and cannot calibrate it. At $p = 1$, total mass alone does not determine the positive count multiset. The proposition concerns exact laws: estimating the maximum of the support from data may require waiting for an event of probability p^r . It supplies no uniform efficient calibration algorithm from finite samples. Known total mass $M > 0$ removes this support-estimation step, since the mean then gives $p = \mathbb{E}Y/M$.

For binary counts, the calibration step is the classical binomial mean–variance identity. Hartung, Holden and Peres [30, Lemma 4.1] use the same identity to estimate unknown symbol-retention probabilities from traces. We state its consequence for aggregate reports and combine it with two-leg recovery.

Corollary 4.4 (Binary calibration from two moments). *For a nonempty binary population on any rooted tree, let $m = \mathbb{E}Y$ and $v = \text{Var}(Y)$. Then*

$$p = 1 - \frac{v}{m}, \quad r = \frac{m}{p},$$

where r is the number of occupied vertices. In particular, for binary two-leg populations and unknown $0 < p < 1$, the mean and covariance of the report determine p and the unordered pair of legs.

Proof. Here $Y \sim \text{Bin}(r, p)$, so $m = rp > 0$ and $v = rp(1-p)$. The report mean μ and covariance C give $m = \sum_k \mu_k$ and $v = \sum_{k,l} C_{kl}$. Recover p , then apply Theorem 5.4. $\square$

Example 4.5 (Unknown survival can confound general counts). On two one-relay legs, the populations $\{0, 2\}$ at $p = 3/8$ and $\{1, 2\}$ at $p = 1/4$ both give

$$\mathbb{E}R(1) = \frac{3}{4}, \quad \text{Var}(R(1)) = \frac{15}{16}.$$

Their laws differ: their maximum possible reports are 2 and 3. Thus joint recovery of arbitrary integer counts and an unknown p from mean and covariance fails even for known geometry and counts bounded by 2. The binary hypothesis in Corollary 4.4 has substantive content.

5 The interior sensor: failures break folded trades

Consider a sensor in the interior of a path, with legs carrying $a, b \in \mathbb{N}^n$. Its noiseless report is the fold $c = a + b$. On equal legs of length n , every report value c is produced by the $\prod_i (c_i + 1)$ splittings $a' + b' = c$, and the invisible trades are exactly the antisymmetric signed populations $(z, -z)$. Without failures, an interior sensor recovers nothing beyond the fold.

Example 5.1. Targets at distances 1 and 2 on the same side, $a = (1, 1)$ and $b = (0, 0)$, are noiselessly indistinguishable from targets at distance 1 on one side and 2 on the other, $a' = (1, 0)$ and $b' = (0, 1)$. With failures they separate. The report puts two targets at hop 1 with probability p^2q for the second population, when the target at distance 2 survives while its inner relay fails, but never for the first, because two relays on one leg are never heard at the same hop.

The mean recovers the fold $c = a + b$, but cannot say how its entries are split between the legs. For $0 < p < 1$, covariance retains information about this split: different legs fail independently, whereas positions on one leg share their surviving ancestors. We will isolate a quadratic measurement of the difference $d = a - b$, locate its deepest nonzero entry, and recover the remaining entries by downward substitution.

5.1 Covariance on one leg

For a leg carrying $u \in \mathbb{N}^n$ write R^u for its report, $P_i(k)$ for the probability that relay i is heard at hop k , and $P_{ij}(k, l)$ for the probability that relay i is heard at hop k and relay j at hop l . Both are explicit: $P_i(k)$ is given by Lemma 2.2, and for $i < j$

$$P_{ij}(k, l) = \begin{cases} P_i(k) \binom{j-i-1}{l-k-1} p^{l-k} q^{(j-i)-(l-k)}, & 1 \leq k \leq i, \quad 1 \leq l-k \leq j-i, \\ 0, & \text{otherwise,} \end{cases}$$

since relay j is heard $l - k$ hops beyond relay i exactly when it survives together with $l - k - 1$ of the $j - i - 1$ relays between them; for $i > j$, $P_{ij}(k, l) = P_{ji}(l, k)$. Then $\mathbb{E}R^u(k) = \sum_i u_i P_i(k)$ and $\mathbb{E}R^u(k)R^u(l) = \sum_{i,j} u_i u_j P_{ij}(k, l)$. We need only five structural identities for the injectivity argument. The closed forms above will also let us evaluate the covariance array in the reconstruction algorithm. Here $[E]$ is 1 when E holds and 0 otherwise:

- (F1) $P_i(k) = 0$ whenever $k > i$, and $P_{ij}(k, l) = 0$ whenever $k > i$ or $l > j$: a relay is never heard further out than its true distance;
- (F2) $P_i(i) = p^i$: relay i is heard at hop i exactly when relays $1, \dots, i$ all survive;
- (F3) $P_{ij}(k, k) = 0$ for $i \neq j$: two surviving relays are never heard at the same hop;
- (F4) $P_{ii}(k, l) = [k = l] P_i(k)$;
- (F5) for $i < m$, $P_{im}(k, m) = [k = i] p^m$: if relay m is heard at hop m , relays $1, \dots, m$ are alive and relay i is heard at hop i .

The same report formula defines R^u for a signed vector $u \in \mathbb{R}^n$. Its covariance array is

$$Q(u)(k, l) = \sum_{i,j} u_i u_j (P_{ij}(k, l) - P_i(k)P_j(l)).$$

Each entry of Q is quadratic in u . Thus $Q(-u) = Q(u)$, and the parallelogram identity gives

$$Q(a + b) + Q(a - b) = 2Q(a) + 2Q(b). \quad (3)$$

The observed covariance is $Q(a) + Q(b)$ by independence. Once the mean has supplied $c = a + b$, this identity therefore makes $Q(a - b) = 2\text{Cov}(R) - Q(c)$ observable. The remaining issue is whether this quadratic array determines its signed input.

For $0 < p < 1$, the last nonzero diagonal entry locates the deepest nonzero coordinate and its magnitude. After one choice of sign, the corresponding column recovers all shallower coordinates. The next lemma gives the identities needed for both steps.

Lemma 5.2 (Deepest entry). *Let $d \in \mathbb{R}^n$ and $m \in \{1, \dots, n\}$ satisfy $d_j = 0$ for all $j > m$. Then*

(a) $Q(d)(k, k) = 0$ for every $k > m$;

(b) $Q(d)(m, m) = d_m^2 p^m (1 - p^m)$;

(c) $Q(d)(k, m) = d_m p^m (d_k - \sum_i d_i P_i(k))$ for every $k < m$.

Proof. For (a), every term with $d_i d_j \neq 0$ has $i, j \leq m < k$. Both the joint landing probability and $P_i(k)$ vanish by (F1), so the diagonal entry is zero.

For (b), only $i = j = m$ can contribute at hop m . Its contribution is $d_m^2 (P_m(m) - P_m(m)^2)$ by (F4), and $P_m(m) = p^m$ by (F2).

For (c), only $j = m$ contributes. When $i < m$, identity (F5) gives $P_{im}(k, m) = [k = i]p^m$; the term $i = m$ has $P_{mm}(k, m) = 0$ because $k < m$. Consequently

$$\begin{aligned} Q(d)(k, m) &= d_k d_m p^m - \sum_i d_i d_m P_i(k) p^m \\ &= d_m p^m \left(d_k - \sum_i d_i P_i(k) \right). \end{aligned}$$

□

5.2 Recovery up to exchange

Proposition 5.3 (Recovery up to sign). *Let $0 < p < 1$ and $d, e \in \mathbb{R}^n$ with $Q(d) = Q(e)$. Then $d = e$ or $d = -e$.*

Proof. First locate the deepest nonzero coordinate. If $d \neq 0$ and m is its largest nonzero index, Lemma 5.2 gives

$$Q(d)(m, m) = d_m^2 p^m (1 - p^m) > 0, \quad Q(d)(k, k) = 0 \quad (k > m).$$

Thus the diagonal vanishes identically exactly when $d = 0$; otherwise its last nonzero position is m . Equality of the arrays implies that either both vectors are zero, or both have the same deepest index m and $d_m^2 = e_m^2$.

In the nonzero case, replace e by $-e$ if necessary so that $e_m = d_m$. This preserves $Q(e)$. For $k < m$, equality of the m th columns and Lemma 5.2(c) allow division by $d_m p^m \neq 0$. Writing $g = d - e$, we obtain

$$(1 - p^k)g_k - \sum_{k < i \leq m} g_i P_i(k) = 0.$$

Since $g_i = 0$ for $i \geq m$ and every pivot $1 - p^k$ is positive, downward substitution gives $g_k = 0$ for all $k < m$. Hence $d = e$ after the possible sign change. □

Theorem 5.4 (Failures break folded trades). *Let $0 < p < 1$. For an interior sensor on a path with legs carrying $a, b \in \mathbb{N}^n$, the mean vector $(\mathbb{E}R(k))_k$ and covariance matrix $(\text{Cov}(R(k), R(l)))_{k,l}$ of the lossy report determine the unordered pair $\{a, b\}$. Equivalently, the mean and second moments determine this pair. In particular the law of the report determines the population up to reflection.*

Proof. Recover $c = a + b$ from the mean using Theorem 4.1. The mean and second moments determine the covariance by $\text{Cov}(R(k), R(l)) = \mathbb{E}R(k)R(l) - \mathbb{E}R(k)\mathbb{E}R(l)$, and conversely.

The legs use disjoint relay failures, so Lemma 2.3 gives $\text{Cov}(R) = Q(a) + Q(b)$. The parallelogram identity (3) then isolates the difference:

$$\Phi := 2 \text{Cov}(R) - Q(c) = Q(a - b).$$

By Proposition 5.3, this determines $d = a - b$ up to sign. Finally,

$$\{a, b\} = \{(c + d)/2, (c - d)/2\}.$$

Changing the sign of d only exchanges the two legs. $\square$

Swapping the independent leg reports preserves their sum's law. Thus, for $0 < p < 1$, two populations have the same lossy law if and only if their padded leg words agree up to exchange. At $p = 1$, every pivot $1 - p^k$ vanishes and $Q \equiv 0$; at $p = 0$, every report vanishes. The nontrivial survival interval is therefore essential for the general identifiability statement.

Proposition 5.5 (Explicit reconstruction). *Let $0 < p < 1$, and let R be the report from legs $a, b \in \mathbb{N}^n$. The following procedure recovers $c = a + b$, a choice of $d = \pm(a - b)$, and the unordered pair of legs. After step 1, set $\Phi = 2 \text{Cov}(R) - Q(c)$.*

1. For $k = n, \dots, 1$: $c_k = (\mathbb{E}R(k) - \sum_{i>k} c_i P_i(k))/p^k$.
2. If $\Phi(k, k) = 0$ for all k , set $d = 0$ and go to step 4. Otherwise let m be the largest k with $\Phi(k, k) \neq 0$, put $d_j = 0$ for $j > m$ and $d_m = \sqrt{\Phi(m, m)/(p^m(1 - p^m))}$.
3. For $k = m - 1, \dots, 1$: $d_k = \left(\frac{\Phi(k, m)}{d_m p^m} + \sum_{k < i \leq m} d_i P_i(k) \right) / (1 - p^k)$.
4. Output $\{(c + d)/2, (c - d)/2\}$.

Proof. This is the proof of Theorem 5.4, with Lemma 5.2 solved explicitly; the choice $d_m > 0$ fixes the reflection. $\square$

The exact and empirical reconstruction procedures are implemented and checked in the accompanying scripts; see Section 8.

5.3 Reconstruction from samples

Integer counts allow each coordinate to be rounded before the next is recovered, so errors do not accumulate. The important scale at hop k is $1 - p^k$: it appears both in the covariance inversion pivot and in the report variance. We retain that scale rather than replace every hop by the smallest gap.

Proposition 5.6 (Variance-sensitive recovery). *Let $n, h \geq 1$ be integers, $0 < p < 1$, and $0 < \delta < 1$. Suppose the two legs have length n and every count is at most h . The empirical estimator specified below recovers $\{a, b\}$ with probability at least $1 - \delta$ from any integer number of independent rounds satisfying*

$$N \geq 2000h^4 p^{-2n} (1 - p)^{-1} \log(4n^2/\delta). \quad (4)$$

For rational p , its exact-arithmetic running time is polynomial in $N, n, \log(h + 1)$ and the bit length of p .

Write $R_k = R(k)$, $\mu_k = \mathbb{E}R_k$, and $C = \text{Cov}(R)$. Given N samples, use their empirical means $\hat{\mu}_k$ and covariance $\hat{C}$ with divisor N . Put $t = p^n$ and $s_k = 1 - p^k$. Perform the following operations, using nearest-integer rounding with any fixed tie convention:

1. Recover the fold downward:

$$\hat{c}_k = \text{round}\left(\frac{\hat{\mu}_k - \sum_{i>k} \hat{c}_i P_i(k)}{p^k}\right), \quad k = n, \dots, 1.$$

2. Set $\hat{\Phi} = 2\hat{C} - Q(\hat{c})$. If no index satisfies $\hat{\Phi}_{kk} > ts_k/2$, put $\hat{d} = 0$. Otherwise let m be the largest such index, put $\hat{d}_i = 0$ for $i > m$, and set

$$v = \text{round}\left(\frac{\hat{\Phi}_{mm}}{p^m s_m}\right), \quad \hat{d}_m = \sqrt{v} > 0.$$

3. In the latter case, recover the remaining differences downward:

$$\hat{d}_k = \text{round}\left(\frac{\hat{\Phi}_{km}/(\hat{d}_m p^m) + \sum_{i>k} \hat{d}_i P_i(k)}{s_k}\right), \quad k = m-1, \dots, 1.$$

4. Output $\{(\hat{c} + \hat{d})/2, (\hat{c} - \hat{d})/2\}$.

If the square root is not a positive integer, or the output fails integrality, nonnegativity or the bound h , return a fixed admissible pair. These fallback choices make the estimator total and are never used on the event proved below. The hop-dependent threshold $ts_k/2$ replaces the common threshold $p^n q/2$ used in version 1. The closed landing formulas of Section 5 evaluate each required entry of Q in polynomial arithmetic time.

Lemma 5.7 (Rare changes at shallow hops). *Under the hypotheses of Proposition 5.6, put $D = 2h$ and $Z_{kl} = (R_k - \mu_k)(R_l - \mu_l)$. Then, for $k \leq l$,*

$$\text{Var}(R_k) \leq 2D^2 s_k, \quad \text{Var}(Z_{kl}) \leq 2D^4 s_k.$$

Moreover $|R_k - \mu_k| \leq D$ and $|Z_{kl} - \mathbb{E}Z_{kl}| \leq 2D^2$.

Proof. The event that the first k relays on both legs survive has probability p^{2k} and forces $R_k = c_k$. Both R_k and c_k lie in $[0, D]$. Consequently

$$\text{Var}(R_k) \leq \mathbb{E}(R_k - c_k)^2 \leq D^2(1 - p^{2k}) \leq 2D^2(1 - p^k).$$

Since $|R_l - \mu_l| \leq D$, we have $\text{Var}(Z_{kl}) \leq \mathbb{E}Z_{kl}^2 \leq D^2 \text{Var}(R_k)$. The almost-sure bounds follow from the same interval constraint. $\square$

We use the classical bounded-variable Bernstein inequality [4]: if independent copies of a centered variable X obey $|X| \leq B$ and $\mathbb{E}X^2 \leq v$, then

$$\mathbb{P}\{|\bar{X}| > \epsilon\} \leq 2 \exp\left(-\frac{N\epsilon^2}{2(v + B\epsilon/3)}\right). \quad (5)$$

For completeness, the usual proof bounds the terms of the exponential series using $\mathbb{E}|X|^j \leq B^{j-2}\mathbb{E}X^2$ and $j! \geq 2 \cdot 3^{j-2}$ for $j \geq 2$. This gives $\log \mathbb{E}e^{\lambda X} \leq \lambda^2 v / [2(1 - \lambda B/3)]$ for $0 \leq \lambda < 3/B$. Independence, exponential Markov, and $\lambda = \epsilon/(v + B\epsilon/3)$ give each tail in (5); the zero-variance case is immediate.

Proof of Proposition 5.6. For analysis only, let $\bar{Z}_{kl}$ be the empirical average of the products centered at the true means, and put $e_k = \hat{\mu}_k - \mu_k$. Define

$$\alpha_k = \frac{ts_k}{5}.$$

Apply (5) to e_k with tolerance α_k and to $\bar{Z}_{kl} - C_{kl}$ with tolerance $\alpha_{\min(k,l)}$. For the product statistic, writing $s = s_{\min(k,l)}$, its denominator obeys

$$2(v + B\alpha/3) \leq 4D^4s + \frac{4}{15}D^2ts \leq 5D^4s.$$

The same upper bound applies to the mean statistic, since $D \geq 2$. Each failure probability is consequently at most

$$2 \exp\left(-\frac{Nt^2s}{125D^4}\right) \leq 2 \exp\left(-\frac{Nt^2q}{2000h^4}\right).$$

There are at most $n + n^2 \leq 2n^2$ statistics. The budget (4) and a union bound give a common event of probability at least $1 - \delta$ on which

$$|e_k| \leq \alpha_k, \quad |\bar{Z}_{kl} - C_{kl}| \leq \alpha_{\min(k,l)}. \quad (6)$$

The estimator does not need the true means to compute its empirical covariance.

Work on this event. Downward fold recovery is exact: once deeper entries are correct, the error in the next unrounded entry is $|e_k|/p^k \leq p^{n-k}s_k/5 < 1/2$. Thus $\hat{c} = c$ and $Q(c)$ is exact. The sample covariance identity is

$$\hat{C}_{kl} = \bar{Z}_{kl} - e_k e_l.$$

For $k \leq l$, since $t, s_l \leq 1$, this implies

$$\begin{aligned} |\hat{\Phi}_{kl} - Q(d)_{kl}| &\leq 2(\alpha_k + \alpha_k \alpha_l) \\ &\leq \frac{12}{25}ts_k < \frac{1}{2}ts_k. \end{aligned}$$

By symmetry the bound holds with $s_{\min(k,l)}$ for all entries. If $d = 0$, every diagonal is below its threshold. Otherwise let m be the last nonzero coordinate of d . Lemma 5.2(a) excludes every $k > m$ from the threshold test, while integrality gives

$$Q(d)_{mm} = d_m^2 p^m s_m \geq ts_m.$$

Therefore the selected index is exactly m . Division by $p^m s_m$ leaves an error at most $12/25 < 1/2$, so rounding recovers d_m^2 and its positive square root fixes the sign of d . At each earlier step the already rounded entries are exact; by Lemma 5.2(c) the only remaining error has magnitude at most

$$\frac{(12/25)ts_k}{|d_m|p^m s_k} \leq \frac{12}{25} < \frac{1}{2}.$$

Downward rounding recovers every coordinate. The output is exactly the true unordered pair, completing the proof. $\square$

Corollary 5.8 (The interior sensor is at least as hard as the endpoint). *For every fixed $q \in (0, 1)$ and fixed positive success probability, recovering from interior-sensor reports the 0/1 populations supported on the first m relays of one leg needs $m^{\Omega(\log \log m)}$ rounds.*

Proof. If $b = 0$, the interior report has the law of the plain endpoint report of a , and the pair $\{a, 0\}$ determines a . Apply Theorem 3.2. $\square$

For 0/1 populations on legs of length n , at fixed $0 < p < 1$ and success probability $2/3$, the minimum round budget for the interior sensor lies between $n^{\Omega(\log \log n)}$ and $\exp(O(n))$. The optimal rate is open (Problem 9.2). For $n \geq 2$, first moments alone cannot identify every two-leg population (Corollary 4.2 and Example 5.1).

5.4 Which covariance entries and which survival probability?

The proof yields two further consequences: exact recovery needs only a small part of the covariance matrix, and the sufficient sample budget behaves differently when survival can be designed.

Corollary 5.9 (A diagonal and one column suffice). *Fix $0 < p < 1$ and $n \geq 1$. The mean vector and covariance diagonal determine $c = a + b$ and whether $a = b$. If $a \neq b$, they also determine the deepest index m at which the legs differ. The additional entries*

$$\text{Cov}(R(k), R(m)), \quad 1 \leq k < m,$$

then determine the unordered pair $\{a, b\}$. Thus exact recovery uses at most $2n - 1$ distinct covariance entries, with the column chosen from the diagonal data.

Proof. Recover c from the mean. The diagonal of $\Phi = 2 \text{Cov}(R) - Q(c) = Q(a - b)$ is then known. By Lemma 5.2, it vanishes identically precisely when $a = b$; otherwise its last nonzero entry locates m and gives $(a_m - b_m)^2$. The indicated observed column entries determine $\Phi(k, m)$, and step 3 of Proposition 5.5 recovers the remaining differences. There are $n + (m - 1) \leq 2n - 1$ observed entries. For $n = 0$ the population is empty without any observations. $\square$

This is an adaptive statement about exact statistics; it does not provide a fixed, population-independent choice of one column. Theorem 5.14 controls the statistical selection errors by estimating the diagonal in a first batch and the selected column in an independent second batch.

Corollary 5.10 (Probability-independent equivalence classes). *For two populations on padded two-leg paths, equality of report laws at one $p \in (0, 1)$ implies equality at every $s \in [0, 1]$. Hence the equivalence classes of populations are the same at all interior survival probabilities.*

Proof. Equality at an interior probability forces the leg pairs to agree up to exchange by Theorem 5.4. Such an exchange preserves the law at every s , including the two endpoints. The converse for equality at every s is immediate. Equality at an endpoint alone need not imply equality in the interior. $\square$

Corollary 5.11 (Designed thinning). *The probability-dependent factor in (4) is uniquely minimized at*

$$p_* = \frac{2n}{2n + 1}.$$

At this probability it suffices to use

$$N = \lceil 2000e h^4 (2n + 1) \log(4n^2/\delta) \rceil. \quad (7)$$

The expected number of inactive relays per round is $2n/(2n + 1) < 1$.

Proof. For $f(p) = p^{-2n}(1-p)^{-1}$,

$$\frac{d}{dp} \log f(p) = \frac{(2n+1)p - 2n}{p(1-p)}.$$

The sign changes from negative to positive at p_* . There $f(p_*) = (2n+1)(1+1/(2n))^{2n} < e(2n+1)$. The failure-count expectation is $2n(1-p_*)$. $\square$

This optimizes the sufficient bound in Proposition 5.6, not the minimax risk over all probabilities or adaptive schedules. The sharper two-batch bound below eliminates the factor $\log n$. Both designed regimes have $p = p_n \rightarrow 1$, so neither contradicts the fixed-probability lower bounds. Controlled thinning presupposes a logical network capable of the stated masking and bypass operation; additional thinning cannot raise a lower baseline survival probability to p_* .

5.5 A lower bound from a folded trade

Theorem 5.12 (Two-point lower bound). *Let $n \geq 2$, $h \geq 1$, $0 < p < 1$ and $0 < \delta < 1/2$. Every possibly randomized estimator that succeeds with probability at least $1 - \delta$ on every two-leg population with legs of length n and counts at most h satisfies*

$$N \geq \frac{\log(1/(2\delta))}{-\log(1 - 2p^2(1-p))}. \quad (8)$$

At $p = p_* = 2n/(2n+1)$, this implies $N \geq n \log(1/(2\delta))$.

Proof. Compare the distinct unordered pairs

$$A = \{(1, 1, 0, \dots), 0\}, \quad B = \{(1, 0, 0, \dots), (0, 1, 0, \dots)\}.$$

This is the folded trade of Example 5.1, now used quantitatively. Trailing zero relays do not affect any occupied relay's report. The only possible nonzero coordinates are the first two, and their exact probabilities are

(R_1, R_2)	$(0, 0)$	$(1, 0)$	$(2, 0)$	$(0, 1)$	$(1, 1)$
P_A	q^2	$2pq$	0	0	p^2
P_B	q^2	$pq(1+q)$	p^2q	p^2q	p^3

The pointwise minimum of the two laws is a subprobability measure of mass $c = 1 - 2p^2q$. Its N -fold product is dominated by both experiment laws and has mass c^N . On that common submeasure the two testing errors sum to its entire mass, even for randomized tests. Hence their sum is at least c^N . An estimator succeeding on both populations gives a test with sum error at most 2δ , so $2\delta \geq c^N$, which is (8). Finally,

$$c - p^2 = (1-p)^2(1+2p) \geq 0.$$

At $p = p_*$ this gives $-\log c \leq 2 \log(1 + 1/(2n)) \leq 1/n$, proving the stated consequence. $\square$

The lower bound is for the specified common probability, not for arbitrary experiments that vary it between rounds. The construction has only two occupied targets; no growing total population is needed for this obstruction.

5.6 Sharp order with linear storage

A uniform union bound discards useful information: the failure probability of the estimate at hop k decreases exponentially in k . The two-batch implementation permits us to sum these probabilities and eliminate $\log n$.

Lemma 5.13 (Hop scales). *For $1 \leq k \leq n$ and $0 < p < 1$,*

$$1 - p^k \geq \frac{k}{n}(1 - p^n).$$

Proof. The sequence $1, p, \dots, p^{n-1}$ is decreasing. Its first k terms have average at least the average of all n terms. Multiply by $k(1 - p)$. $\square$

Theorem 5.14 (Two-batch recovery). *Under the assumptions of Proposition 5.6, put $t = p^n$ and use two independent batches, each of size*

$$N_0 = \left\lceil \frac{2000h^4n}{t^2(1-t)} \log(1 + 8/\delta) \right\rceil. \quad (9)$$

There is an estimator with success probability at least $1 - \delta$ that retains $O(n)$ scalar accumulators between rounds.

Proof. The first batch stores n sums and n sums of squares. Its means and diagonal covariance entries determine c , m , and d_m by the preceding rounding proof (or determine that $d = 0$). For each k , the two analyzed statistics have combined failure probability at most

$$4 \exp\left(-\frac{N_0 t^2 s_k}{2000h^4}\right).$$

By Lemma 5.13, writing $A = N_0 t^2(1 - t)/(2000h^4n)$, their total failure probability is at most

$$4 \sum_{k=1}^n e^{-Ak} \leq \frac{4}{e^A - 1} \leq \frac{\delta}{2}.$$

If $d \neq 0$, the second batch stores its own sums for coordinates $1, \dots, m$ and the products of each with coordinate m . On the successful first-batch event the selected m is the true one. Independence permits the same concentration argument conditionally on the first batch. Each coordinate $k \leq m$ again has at most two analyzed statistics, with error bounds proportional to s_k ; the mean at m supplies the factor needed to center the whole covariance column. The same geometric sum bounds the second failure probability by $\delta/2$. Use the exact recovered c and d_m with the second batch's covariance column in the downward recurrence. The rounding proof applies without change.

Evaluate only the required diagonal or column entries of $Q(c)$, on demand, rather than storing its entire matrix. The algorithm needs $O(n)$ scalar accumulators. An invalid first-batch result can cause immediate fallback; its probability is already included in the first failure bound. $\square$

Corollary 5.15 (Matching minimax order at a rational design). *Let $n \geq 2$, $h \geq 1$, $0 < \delta \leq 1/4$, and $p_n = 2n/(2n+1)$. Write $N^*(n, h, \delta; p_n)$ for the smallest fixed worst-case number of independent reports allowing success at least $1 - \delta$ uniformly over $a, b \in \{0, \dots, h\}^n$. Then*

$$n \log(1/(2\delta)) \leq N^*(n, h, \delta; p_n) \leq 2 \lceil 16000h^4n \log(1 + 8/\delta) \rceil. \quad (10)$$

In particular $N^*(n, h, \delta; p_n) = \Theta_h(n \log(1/\delta))$ for each fixed h . The upper bound is achieved with $O(n)$ scalar accumulators.

Proof. Bernoulli's inequality gives

$$p_n^n = \left(1 - \frac{1}{2n+1}\right)^n \geq 1 - \frac{n}{2n+1} > \frac{1}{2}, \quad p_n^{-n} = \left(1 + \frac{1}{2n}\right)^n \geq \frac{3}{2}.$$

Hence $1/2 \leq t = p_n^n \leq 2/3$. The function $t^2(1-t)$ increases on $[1/2, 2/3]$, so $t^2(1-t) \geq 1/8$. Substitute in (9) and use Theorem 5.12. On $0 < \delta \leq 1/4$, both logarithmic expressions are comparable to $\log(1/\delta)$ with absolute constants. $\square$

The storage claim counts scalar accumulators, not bits. It does not determine the minimum number of fixed covariance entries (Problem 9.6). The power h^4 is a sufficient dependence, with no matching claim in h .

For the particular bound (9), the ideal real-valued design is $p^n = 2/3$, since $t^2(1-t)$ is maximized there. It gives batch coefficient $13500h^4n$. The rational choice in Corollary 5.15 is within a constant factor and supports the exact rational implementation directly. Neither optimization establishes optimality over all possible measurement schedules or communication protocols.

6 Spiders, trees and a Thue–Morse obstruction

A *spider* is a tree whose root has r children, each starting a path (a leg). An interior path sensor is a spider with two legs. Write a population on a spider with legs of length L as the multiset $\{w_1, \dots, w_r\}$ of leg words; by Proposition 2.4 the law depends only on this multiset.

6.1 The exact parity obstruction

For $0 < p < 1$, the means and covariances determine the two leg words up to exchange. With more legs, independence still makes cumulants additive, but parity cancellation can conceal different populations through any prescribed order. The moment–cumulant relations and additivity under independent vector sums are standard [33, Sections 2.3 and 2.7.1]. We use the classical even/odd Boolean trade [20, 21]; deterministic distance-moment compression and tensor propagation using this cancellation appear in [2, Theorems 21 and 24]. Its application below concerns stochastic joint moments of superimposed leg reports, and identifies their first unequal moment explicitly.

Theorem 6.1 (Thue–Morse spiders). *Let $k \geq 1$ and consider the spider with 2^k legs of length $k+1$. Let x^{even} place the 2^k words of $\{0,1\}^{k+1}$ with an even number of ones on the legs, and x^{odd} the 2^k words with an odd number. For every $p \in (0,1]$, the lossy reports of x^{even} and x^{odd} have the same joint moments $\mathbb{E} \prod_{t=1}^j R(l_t)$ for every $j \leq k$ and all hops $l_1, \dots, l_j$. In particular they have the same noiseless report, and for $k \geq 2$ the same mean and covariance. Writing R_e and R_o for the even and odd reports, respectively, their first-hop moments satisfy*

$$\mathbb{E}R_e(1)^{k+1} - \mathbb{E}R_o(1)^{k+1} = -k!(k+1)!p^{k+1}q^{k(k+1)/2}. \quad (11)$$

Thus, for $0 < p < 1$, their laws differ and the first order at which a joint moment differs is exactly $k+1$.

Proof. Equality through order k . For bounded random variables $Y_1, \dots, Y_j$, their joint cumulant is the derivative $\partial_{t_1} \cdots \partial_{t_j} \log \mathbb{E} \exp(\sum_s t_s Y_s)$ at $t = 0$. Expanding the logarithm expresses it as a sum of products of moments, each with total order j . Independence turns the moment-generating function of a sum into a product, so joint cumulants add over independent legs.

For a leg word $w \in \{0, 1\}^{k+1}$, each report entry $R^w(l)$ is linear in w . Every joint moment of order j , and hence every joint cumulant of that order, is therefore a polynomial in w of degree at most j . The polynomial is the same on every leg because the lengths and survival probabilities agree. On the binary cube, reduce $w_i^2 = w_i$. If $j \leq k$, each remaining monomial omits at least one coordinate, and

$$\sum_{w \in \{0,1\}^{k+1}} (-1)^{|w|} \prod_{i \in S} w_i = 0 \quad (|S| \leq k).$$

Indeed, summing first over a coordinate outside S cancels its two terms. Additivity now gives equal cumulants through order k for the two full reports. Exponentiating their log moment-generating functions gives equal joint moments through the same order.

The first unequal moment. Put $\ell = k + 1$. The first-hop report of one binary leg is Bernoulli with parameter

$$z(w) = \sum_{i=1}^{\ell} p q^{i-1} w_i :$$

relay i is the first survivor exactly when it survives and its $i-1$ predecessors fail. By independence, the first-hop moment-generating functions are

$$M_{\epsilon}(t) = \prod_{|w| \equiv \epsilon \pmod{2}} (1 + z(w)(e^t - 1)), \quad \epsilon \in \{0, 1\}.$$

Let $S_r = \sum_w (-1)^{|w|} z(w)^r$. The same cancellation gives $S_r = 0$ for $r < \ell$. At degree ℓ , only the monomial containing every coordinate survives; its coefficient in $z(w)^{\ell}$ is $\ell! \prod_{i=1}^{\ell} p q^{i-1}$. Hence

$$S_{\ell} = (-1)^{\ell} \ell! p^{\ell} q^{\ell(\ell-1)/2}.$$

Expanding the logarithms as formal power series at zero yields

$$\begin{aligned} \log M_0(t) - \log M_1(t) &= \sum_{r \geq 1} \frac{(-1)^{r-1}}{r} S_r (e^t - 1)^r \\ &= -(\ell - 1)! p^{\ell} q^{\ell(\ell-1)/2} t^{\ell} + O(t^{\ell+1}). \end{aligned}$$

Both generating functions have constant term 1, so the same leading-order expansion holds for $M_0 - M_1$. The coefficient of t^{ℓ} in $M_0 - M_1$ is the difference of the ℓ th raw moments divided by $\ell!$, proving (11). Its right-hand side is negative for $0 < p < 1$. $\square$

The distinction between the full laws also has a direct explanation. The even population has an empty leg, so $R_e(1) \leq 2^k - 1$. Every odd leg is nonempty, and for $0 < p < 1$ each has a positive chance $z(w)$ of contributing a target at hop 1. Independence gives

$$\mathbb{P}\{R_o(1) = 2^k\} = \prod_{|w| \text{ odd}} z(w) > 0.$$

Thus the example hides information from low-order statistics without destroying it in the full law. No fixed moment order identifies populations on all spiders. For $k = 1$ the pair is $\{00, 11\}$ against $\{01, 10\}$, a folded trade on two legs, and Theorem 5.4 shows that order two suffices there. For $k = 2$, at $p = 1/2$, it is the only coincidence of means and covariances among all 330 binary populations on the four-legged spider with legs of length 3, while on three legs of length 3 or

Family	Populations	Instances	Lossy classes	Noiseless classes
Endpoint path, $n = 6$	0/1	64	64	64
Two legs, $L = 3$	0/1	36	36	27
Two legs, $L = 4$	0/1	136	136	81
Two legs, $L = 5$	0/1	528	528	243
Two legs, $L = 3$	counts ≤ 2	378	378	125
Three legs, $L = 3$	0/1	120	120	64
Three legs, $L = 4$	0/1	816	816	256
Four legs, $L = 3$	0/1	330	330	125
Rooted trees, 1–8 non-root vertices (485)	0/1	91,834	no collision	–
Rooted trees, 1–6 non-root vertices, $p = \frac{1}{3}$	0/1	3,898	no collision	–
Rooted trees, 1–5 non-root vertices	counts ≤ 2	5,718	no collision	–

Table 1: Exact law comparisons. For spiders, instances are populations up to permutations of legs; “lossy classes” counts distinct report laws at $p = 1/2$ (identical at $p = 1/3$), “noiseless classes” distinct noiseless reports. For trees, all (tree, population) instances on all trees up to the stated size were compared with each other, with reports compared after removing trailing zeros; equal laws occur only for isomorphic occupied cores (6,941 distinct laws in the first tree row). Unless stated, $p = 1/2$. The tree rows with $p = 1/2$ are extended to every $p \in (0, 1)$ in the text after Conjecture 6.2.

4 means and covariances separate all 120, respectively 816, populations. Empty legs contribute nothing to the report, and empty relays beyond the occupied ones change nothing, so every spider with at least 2^k legs of length at least $k + 1$ contains this pair: the order of statistics needed to read a spider hub grows at least logarithmically with the number of legs.

The quantitative feature specific to the present model is (11): low-order cancellation coexists with separation of the full lossy laws for every $0 < p < 1$. This proves a lower bound on the moment order needed at a spider hub; it does not assert an exact population capacity analogous to the earlier Hamming-product theorem. Prouhet–Thue–Morse cancellation also underlies the imported endpoint hard pairs [5].

6.2 The tree conjecture

The parity examples separate the amount of information in a law from the order of moments needed to access it. For a general rooted tree, we ask whether the labelled occupied core accounts for every ambiguity in the full law.

Conjecture 6.2 (Complete invariant). *For every $p \in (0, 1)$, the laws $\mu_{T,x,p}$ and $\mu_{T',x',p}$ coincide if and only if the occupied cores of (T, x) and (T', x') are isomorphic as rooted $\mathbb{N}$ -labelled trees.*

The “if” direction is Proposition 2.4. Theorem 5.4 proves the conjecture for paths, with the sensor at any vertex. Table 1 records exact comparisons at fixed p . To cover every p , note that each law is a vector of polynomials in p with integer coefficients, and that by Theorem 4.1 only instances with equal noiseless reports can have equal laws. For each pair of non-isomorphic occupied cores with equal noiseless reports we computed the gcd of all coordinatewise differences of their laws, removed the factors p and $1 - p$, and counted its roots in $(0, 1)$ by Sturm’s theorem. There are none for the 43,619 such pairs among the 6,941 occupied cores of 0/1 populations on trees with one to eight non-root vertices, or for the 4,051 pairs among the 1,536 cores with counts at most 2 on trees with one to five. So Conjecture 6.2 holds for these instances at every $p \in (0, 1)$.

6.3 A recursive description of the full law

The general-tree conjecture can also be phrased as an algebraic uniqueness question. The deterministic shell polynomial used for reliable aggregation in [2, Section 11] records one count at each distance. The polynomial here instead has a coefficient for each possible random report vector, equal to that report's probability. Let $\mathbf{z} = (z_1, z_2, \dots)$ and let $S\mathbf{z} = (z_2, z_3, \dots)$; only finitely many variables will occur. For a non-root vertex v , let $H_v(\mathbf{z})$ be the probability generating polynomial of the report of its entire descendant subtree, with hops measured relative to the parent of v . Write $\mathcal{C}(v)$ for its children.

Proposition 6.3 (Recursive report polynomials). *The branch polynomials satisfy*

$$H_v(\mathbf{z}) = q \prod_{w \in \mathcal{C}(v)} H_w(\mathbf{z}) + pz_1^{x(v)} \prod_{w \in \mathcal{C}(v)} H_w(S\mathbf{z}),$$

with empty products equal to 1. The full report generating polynomial is

$$G_{T,x,p}(\mathbf{z}) = \mathbb{E} \prod_{k \geq 1} z_k^{R_A(k)} = \prod_{v \in \mathcal{C}(\rho)} H_v(\mathbf{z}).$$

In particular, the first-hop branch polynomial $F_v(t) = H_v(t, 1, 1, \dots)$ obeys

$$F_v(t) = pt^{x(v)} + q \prod_{w \in \mathcal{C}(v)} F_w(t).$$

Proof. Condition on v surviving. If it fails, it contributes no count and every child subtree is measured at the same hop positions after bypass, giving the first product. If it survives, its own count lies at hop 1 and every child contribution is shifted outward by one hop, giving the second product. Child subtrees are independent conditional on the status of v . At the sensor, independence of the root branches gives the formula for G . Finally $H_w(1, 1, \dots) = 1$, which gives the first-hop specialization. $\square$

An unoccupied leaf has $H_v = q + p = 1$, and repeated deletion of such leaves leaves G unchanged, recovering the occupied-core reduction. The coefficients of G are the exact report probabilities, so equality of laws is exactly equality of these polynomials. Conjecture 6.2 asks whether this recursive product-and-shift construction has any ambiguity beyond rooted labelled-core isomorphism. The recursion is a way to compute and study the law; no polynomial bound on the size of its expanded coefficient table is claimed. It also clarifies why a first-hop distribution alone can discard information retained by the full hop vector.

7 Formalization

The accompanying Lean 4 [24] project (toolchain `v4.35.0-rc2`, with Mathlib [23] at the matching tag) contains an inherited exact-recovery library with populations $\text{Fin } n \rightarrow \mathbb{N}$; only probability weights, moments and signed differences take values in a linearly ordered field K , such as $\mathbb{R}$ or $\mathbb{Q}$. All compared legs have the same length n . The padding interpretation in Section 2 is not a separate formal theorem between different lengths. Positions and hops are indexed from 0. Reports, laws and moments are finite sums over failure patterns $\omega: \text{Fin } n \rightarrow \text{Bool}$ weighted by $\prod_i (p \text{ if } \omega_i \text{ else } 1 - p)$.

Declaration	Statement
<code>endpoint_report_eq_trace</code>	entry k of the endpoint report is entry k of the trace (Proposition 3.1)
<code>trace_eq_horizon_report</code>	the trace is the first $ A $ report entries (Proposition 3.1)
<code>twoLegReport_noiseless</code>	without failures the interior sensor reports $a + b$
<code>twoLegLaw_one</code>	at $p = 1$ the law is the point mass at $a + b$
<code>twoLegMean_eq_of_fold_eq</code>	equal folds give equal means, for every p (Theorem 4.1)
<code>fold_eq_of_twoLegMean_eq</code>	equal means give equal folds, for $p > 0$ (Theorem 4.1)
<code>legs_eq_of_moments</code>	for $0 < p < 1$, means and second moments determine $\{a, b\}$ (Theorem 5.4)
<code>legs_eq_of_mean_covariance</code>	for $0 < p < 1$, means and covariance determine $\{a, b\}$ (Theorem 5.4)
<code>legs_eq_of_twoLegLaw</code>	for $0 < p < 1$, the law determines $\{a, b\}$ (Theorem 5.4)
<code>twoLegLaw_eq_iff</code>	for $0 < p < 1$, two laws are equal exactly when the leg pairs agree up to exchange
<code>failures_break_trade</code>	Example 5.1: equal noiseless reports and $p = 1$ laws, different laws for $0 < p < 1$

Covariance is the second moment minus the product of means. Exchange symmetry follows by swapping the independent failure patterns. Supporting lemmas give the zero-report point mass and vanishing mean and covariance at $p = 0$. Sparse recovery (Corollary 5.9) first determines the difference up to sign from the diagonal of Q and one selected column, then lifts this to observed means and covariances; full-matrix recovery follows. Further library results formalize Corollary 5.10. These supplement the eleven selected declarations.

Solution.lean exposes the eleven selected proved results. The proof sources and pinned build configuration are included in the supplemental materials. The checked transitive axiom dependencies of 23 exact-recovery endpoints and 76 declarations supporting the sampling analysis are limited to `propext`, `Classical.choice` and `Quot.sound`. Independent human review has not been performed.

The new modules have the following scope:

- **FiniteSample.lean** proves that a surviving prefix fixes the report entry, gives actual-model variance and centered-product bounds, and proves the empirical covariance identity and the weighted error margin.
- **Pipeline.lean** proves recovery of the fold, deepest index, positive square root, and covariance-column solution under explicit accuracy and rounding equations. The nonzero and symmetric cases are covered. It specifies correctness of the reconstruction equations, rather than a formal executable implementation of the total sample estimator.
- **LowerBound.lean** proves the exact length-two trade laws, their total variation and lower bounds for arbitrary randomized tests on any number of independent rounds, including the designed-probability logarithmic bound. Separately, **AllLengthLower.lean** proves directly on actual paths of every length $n \geq 2$ the bound $N \geq (n/2) \log(1/(2\delta))$ at p_n . Its common-prefix event gives a factor-two weaker constant than Theorem 5.12; the sharper constant's padding transfer remains in the written proof.
- **Geometric.lean** and **TailBudget.lean** prove the prefix-scale inequality, rational design bounds, and geometric summation of coordinate-dependent errors.

The new declarations likewise have only the standard axiom dependencies listed above. Bernstein concentration, the probability of the simultaneous good events, conditional independence

for sample splitting, and the final minimax composition remain in the written proof. Thus Proposition 5.6, Theorem 5.14, and Corollary 5.15 are not complete Lean probability theorems.

Also not formalized are the imported OpenAI theorems and their simulation transfer (Theorems 3.2 and 3.3 and Corollary 5.8), the closed landing formula of Lemma 2.2, the general-tree claims of Lemma 2.3, Proposition 2.4, and Theorem 4.1, the spider obstruction Theorem 6.1, general-tree kernels and recursion (Propositions 2.5 and 6.3), calibration (Proposition 4.3 and Corollary 4.4), and the computations of Section 8. The derivative-based optimization of Corollary 5.11 is written only; the rational design inequalities used in Corollary 5.15 are checked separately as described above.

8 Computations

All probability and moment calculations use exact rational arithmetic. The supplemental materials provide the two-batch decoder and exact verification scripts for the version 2 sampling results. The earlier path, spider and tree computations summarized below are documented with the October 10 version [3]. The law comparisons in Table 1 enumerate every failure pattern for each tree and population. Rooted trees are generated up to isomorphism; reports are compared after deleting trailing zeros. Comparing a digest of each exact law with a canonical occupied-core description also checks populations on different trees and at different sizes.

We checked Proposition 5.5 against the exact laws: means and second moments computed by enumerating all failure patterns, for all unordered pairs of legs of length 3 and 4 with counts at most 2 and of length 4 and 5 with counts at most 1, at $p \in \{1/3, 1/2, 3/4\}$. Every pair is recovered exactly (4,227 distinct pairs per value of p). The closed forms of P_i and P_{ij} used by the algorithm agree with exact enumeration for every relay pair and hop pair on legs of length at most 6, at $p \in \{1/3, 1/2, 3/5, 1\}$.

The version 1 empirical implementation uses the closed landing formulas to evaluate covariance, rather than enumerating failure patterns. It recovers all 1,695 exact small test pairs and all 8,640 tested corners of a conservative raw-moment error box with radius $p^n(1-p)/(10(2h+1))$. For the version 1 uniform-threshold concentration argument, a further 8,640 exhaustive centered-event corners and 3,072 stress corners with count bounds up to 100 verify both the exact covariance-error identity and recovery without fallback. These inputs perturb the means and the true-centered second products by at most $\alpha = p^n(1-p)/5$, then convert them to the raw moments received by the unchanged estimator. The raw-moment boxes are stricter subsets of the centered event. Further checks exercise invalid-output fallback and verify $R(k) \leq 2h$. Synthetic corners need not be realizable samples; they test the deterministic implication from the proof’s accuracy event. The success-probability guarantee is supplied by concentration, not by Monte Carlo simulation.

A structural-extension script checks 375 total-count laws for calibration and positive-label reconstruction, and 543 nonempty binary two-leg cases for calibration followed by recovery. It verifies the counterexample of Example 4.5, compares 3,445 joint landing kernels with direct failure-pattern enumeration, and checks the full and first-hop polynomial recursions on 4,290 labelled-tree instances. The associated interior-probability cases also check 2,526 nonempty tree inventories, while 570 insertions of an unoccupied leaf preserve the law. The version 1 design checks compare its quadratic bound on exact rational grids and by finite exponential-series inequalities. These finite checks supplement the general proofs in Propositions 2.5, 4.3 and 6.3 and Corollary 4.4.

The version 2 sampling companion separately checks 177 exact finite experiments and 2,832 admissible perturbation decodes, including the rare-change variance bounds and two-point report

laws. A report-only two-batch implementation retains linear scalar storage; 76 exact empirical-distribution fixtures verify its reconstruction. These fixtures are prescribed multisets of reports, not an independent-sampling experiment. A further exact-rational check verifies 122,500 design, prefix-scale, denominator, tail-budget and overlap inequalities on its stated finite grid. The scripts record their case ranges and results. They support the written proof and do not replace its universal concentration or minimax argument.

The original full-law spider calculations check $k \leq 3$ at $p \in \{1/2, 1/3\}$. A separate Poisson-binomial computation checks the exact moment gap (11), all lower first-hop moments, and the distinguishing support event for $k = 1, \dots, 6$ at $p \in \{1/3, 1/2, 3/4\}$, together with the boundary cases $p = 0, 1$. The arbitrary- k conclusion is proved in Theorem 6.1; it does not rest on these finite checks.

For the general-tree conjecture, integer-polynomial laws and exact gcd/Sturm calculations give the all- p verification described after Conjecture 6.2. These finite checks do not establish the unrestricted tree conjecture.

9 Prospective applications and open problems

9.1 What the model could support

Spatial counts from anonymous telemetry. A logical collection tree may attach a count $x(v)$ to each relay while reporting only counts by hop. At an interior path sensor, repeated covariance measurements can distinguish populations with the same noiseless fold, using no target identifiers or additional sensor positions. If the population is stable across independent rounds and the bypass rule is implemented, this could support occupancy or inventory diagnostics up to the stated reflection. The total-count law provides a separate geometry-free check of the inventory and failure rate. Ordinary collection routing [19, 25] can lose an entire descendant subtree when a relay fails; the theorem requires those descendants to remain reachable through the specified contraction.

Deliberate measurement design. A programmable overlay could choose independent active-relay masks and route through nearest active ancestors. Corollary 5.15 provides a concrete design rule: predominantly active relays, with fewer than one inactive relay per round in expectation, break the fold ambiguity in $\Theta_h(n \log(1/\delta))$ rounds at $p_n = 2n/(2n+1)$, with linear scalar storage. The logical inactive state must silence that relay's own targets while permitting bypass. The bound is a theoretical guarantee for that experiment, not a measured deployment result; routing cost and communication overhead remain to be included in a protocol-level optimization.

Information disclosed by aggregate reports. A noiseless trade can cease to hide a population once repeated relay masks are observed through their aggregate reports. Thus a telemetry audit should distinguish release of a mean histogram, a mean together with covariance, and individual round reports: the first retains exactly the noiseless information, the second identifies two legs, and on spiders a prescribed finite moment order can still miss distinctions present in the full law. These conclusions concern the information content of the stipulated measurements; they neither recover target identities nor establish a differential-privacy guarantee.

9.2 Precise mathematical questions

Problem 9.1 (Fixed-probability endpoint complexity). Determine the optimal round complexity for binary plain endpoint reports at fixed rational $p \in (0, 1)$. The present bounds range from $n^{\Omega(\log \log n)}$ to $\exp(O((\log n)^3(1 + \log \log n)^6))$. Does the absence of the surviving horizon change the minimax rate, despite the common upper bound obtained from the observables of [6]?

Problem 9.2 (Interior complexity and design). Determine the minimax round budget for two-leg populations as a function of n, h, p and confidence. At fixed p , is it quasi-polynomial, or does superposition force $\exp(n^{\Omega(1)})$ rounds? Corollary 5.15 settles the order in n and confidence at $p_n = 2n/(2n + 1)$ for fixed h , but the optimal dependence on h remains unknown. Determine the best common survival probability, or a schedule of known probabilities, under a budget on the expected number of inactive relays. The designs of Corollaries 5.11 and 5.15 optimize sufficient bounds, not all possible experiments.

Problem 9.3 (The occupied-core conjecture). Prove or refute Conjecture 6.2. Can the recursive polynomials of Proposition 6.3 be factored uniquely into labelled rooted branches at every fixed $0 < p < 1$? A counterexample would need equal full generating polynomials for non-isomorphic occupied cores, rather than agreement of finitely many low-order moments.

Problem 9.4 (Moment order on a fixed number of branches). For spiders with r legs, does a finite order $j(r)$, independent of leg lengths and of the known probability $0 < p < 1$, determine every integer population up to leg exchange and trailing-zero padding? If so, find its least value. Theorems 5.4 and 6.1 give $j(2) = 2$ and $j(2^k) \geq k + 1$. The existence of $j(r)$ is not supplied by the fact that each individual finite-support law is determined by all its moments.

Problem 9.5 (Finite-sample calibration). For nonempty two-leg populations of known length n and count bound h , and p unknown in a specified compact subinterval of $(0, 1)$, find efficient quantitative algorithms that recover the population up to exchange and estimate p to prescribed accuracy without relying on observation of the maximum total count. Determine the least sufficient moment order as a function of n and h . Bare existence of a finite-order route is immediate: the first $2nh$ moments of $Y \in \{0, \dots, 2nh\}$, together with its zeroth moment 1, determine its law by Vandermonde inversion; Proposition 4.3 then gives p , and the report mean and covariance recover the legs. This argument supplies no useful conditioning bound. Example 4.5 rules out a universal two-moment answer for general integer counts, whereas binary populations admit Corollary 4.4. Quantify the sample cost in terms of n, h , the prescribed accuracy and confidence, and lower bounds for p and $1 - p$.

Problem 9.6 (Adaptive versus fixed covariance measurements). For fixed known $0 < p < 1$, how many fixed coordinate entries of the covariance matrix, together with the full mean, suffice uniformly over two-leg populations of length n ? Corollary 5.9 gives $2n - 1$ entries with an adaptive column choice. Theorem 5.14 and Corollary 5.15 give round guarantees with linear scalar storage using two independent batches. Determine the optimal adaptive and nonadaptive covariance counts, their required bit storage, and the round cost for a fixed population-independent selection of entries.

Problem 9.7 (Unequal relay reliability). Suppose each relay v has a known rational survival probability $p_v \in [\varepsilon, 1 - \varepsilon]$, independently of the others. Characterize recovery from first and second moments modulo symmetries preserving both the root and the rates. Give quantitative separation bounds in terms of the rate encoding length. Unequal rates can break exchange symmetries, but

arithmetic uniqueness alone need not imply stable recovery. What changes when the rates are only approximately known?

Problem 9.8 (Several sensors and imperfect observations). For a fixed known tree and bounded populations, determine sensor sets whose separate rooted experiments identify the population from first and second moments, modulo automorphisms fixing the sensors. Distinguish those data from synchronized joint sensor reports, which can carry extra cross-sensor information. Then quantify how much dependence between relay failures, hop misclassification, or contamination of the report law can be tolerated. On graphs with cycles, specify the routing rule before seeking analogous results, since different rerouting rules define different inverse problems.

Acknowledgments and responsibility

Anthropic Claude (Opus) and OpenAI ChatGPT/Codex provided substantial assistance with literature review, model and proof development, exact computations, Lean formalization and writing. OpenAI Codex additionally assisted the variance-sensitive sampling analysis, lower bound, implementation and formalization added in version 2. The exposition also benefited from Scott Armstrong and Amélie Loher’s [math-paper-skills](#). The OpenAI preprints [5–7] are used through their stated theorems; their proofs have not been independently verified in preparing this revision. Automated review and proof checking do not constitute independent human peer review. Lennart Rudolph is the sole named author and responsible maintainer; no AI system is an author.

Data and code availability

Version 2.0.0 is distributed as a manuscript PDF and a separate supplemental-materials ZIP. The ZIP contains the Lean proof sources and pinned build files in `lean/`, the decoder and three exact verification scripts in `scripts/`, and their licenses. From `lean/`, run `lake exe cache get` followed by `lake build` under the pinned toolchain. With Python 3.9 or later, run `check_sampling.py`, `check_streaming.py` and `check_design_bounds.py` from `scripts/`; they require only the standard library and the included modules and print their check results. The all-versions DOI is [doi:10.5281/zenodo.23252713](https://doi.org/10.5281/zenodo.23252713). The original document of October 10, 2026 is preserved as version 1.0.0 at [doi:10.5281/zenodo.23252714](https://doi.org/10.5281/zenodo.23252714). The OpenAI PDF links in the bibliography identify the exact cited files at repository commit [adc7f1241b42](https://github.com/anthropics/anthropic-cockpit/commit/adc7f1241b42).

References

- [1] L. Rudolph. Distance-shell tomography on graphs: integer trades and optimal grid sensing. Preprint, Zenodo, 2026. [doi:10.5281/zenodo.22404456](https://doi.org/10.5281/zenodo.22404456).
- [2] L. Rudolph. Distance-shell tomography: geometric certificates for fault-tolerant sensing, moment compression, and hypercube equalization. Preprint, Zenodo, 2026. [doi:10.5281/zenodo.23125523](https://doi.org/10.5281/zenodo.23125523).
- [3] L. Rudolph. Lossy distance-shell tomography on paths: recovery from mean and covariance. Previous version, Zenodo, October 10, 2026. [doi:10.5281/zenodo.23252714](https://doi.org/10.5281/zenodo.23252714).
- [4] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013, Section 2.8. [doi:10.1093/acprof:oso/9780199535255.001.0001](https://doi.org/10.1093/acprof:oso/9780199535255.001.0001).
- [5] OpenAI. Quantitative lower bounds for trace reconstruction. OpenAI Math Release preprint, September 24, 2026. github.com/openai/math.

- [6] OpenAI. Uniform quasipolynomial-time trace reconstruction. OpenAI Math Release preprint, October 5, 2026. github.com/openai/math.
- [7] OpenAI. A latest-anchor induction with spectrally compact masks for worst-case trace reconstruction. OpenAI Math Release preprint, October 5, 2026. github.com/openai/math.
- [8] T. Batu, S. Kannan, S. Khanna, and A. McGregor. Reconstructing strings from random traces. In *Proceedings of the 15th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 910–918, 2004.
- [9] T. Holenstein, M. Mitzenmacher, R. Panigrahy, and U. Wieder. Trace reconstruction with constant deletion probability and related results. In *Proceedings of the 19th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 389–398, 2008.
- [10] A. De, R. O’Donnell, and R. A. Servedio. Optimal mean-based algorithms for trace reconstruction. In *Proceedings of the 49th ACM Symposium on Theory of Computing (STOC)*, pages 1047–1056, 2017. Journal version: *Annals of Applied Probability*, 29(2):851–874, 2019.
- [11] F. Nazarov and Y. Peres. Trace reconstruction with $\exp(O(n^{1/3}))$ samples. In *Proceedings of the 49th ACM Symposium on Theory of Computing (STOC)*, pages 1042–1046, 2017.
- [12] Z. Chase. Separating words and trace reconstruction. In *Proceedings of the 53rd ACM Symposium on Theory of Computing (STOC)*, pages 21–31, 2021.
- [13] A. Burudgunte, P. Valiant, and H. Wang. Quasipolynomial trace reconstruction. Preprint, 2026. arXiv:2607.04073v1.
- [14] N. Holden and R. Lyons. Lower bounds for trace reconstruction. *Annals of Applied Probability*, 30(2):503–525, 2020.
- [15] Z. Chase. New lower bounds for trace reconstruction. *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, 57(2):627–643, 2021.
- [16] S. Davies, M. Z. Rácz, and C. Rashtchian. Reconstructing trees from traces. In *Proceedings of the 32nd Conference on Learning Theory (COLT), Proceedings of Machine Learning Research*, 99:961–978, 2019. [Publisher record and open-access paper](#).
- [17] F. Ban, X. Chen, A. Freilich, R. A. Servedio, and S. Sinha. Beyond trace reconstruction: population recovery from the deletion channel. In *Proceedings of the 60th IEEE Symposium on Foundations of Computer Science (FOCS)*, 2019. arXiv:1904.05532.
- [18] D. Niculescu and B. Nath. DV based positioning in ad hoc networks. *Telecommunication Systems*, 22:267–280, 2003.
- [19] O. Gnawali, R. Fonseca, K. Jamieson, D. Moss, and P. Levis. Collection tree protocol. In *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 2009.
- [20] J.-P. Allouche and J. Shallit. The ubiquitous Prouhet–Thue–Morse sequence. In *Sequences and their Applications*, pages 1–16, 1999. [doi:10.1007/978-1-4471-0551-0_1](https://doi.org/10.1007/978-1-4471-0551-0_1).
- [21] P. Frankl and J. Pach. On the number of sets in a null t -design. *European Journal of Combinatorics*, 4(1):21–23, 1983. [doi:10.1016/S0195-6698\(83\)80004-3](https://doi.org/10.1016/S0195-6698(83)80004-3).
- [22] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.
- [23] The mathlib Community. The Lean mathematical library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP)*, 2020.
- [24] L. de Moura and S. Ullrich. The Lean 4 theorem prover and programming language. In *Automated Deduction – CADE 28*, 2021.
- [25] S. Madden, M. J. Franklin, J. M. Hellerstein, and W. Hong. TAG: a Tiny AGgregation Service for Ad-Hoc Sensor Networks. In *Proceedings of the 5th Symposium on Operating Systems Design and Implementation (OSDI)*, USENIX, 2002. [Publisher record](#).
- [26] N. G. Duffield and F. Lo Presti. Network tomography from measured end-to-end delay covariance. *IEEE/ACM Transactions on Networking*, 12(6):978–992, 2004. [doi:10.1109/TNET.2004.838612](https://doi.org/10.1109/TNET.2004.838612).

- [27] Y. Vardi. Network tomography: Estimating source-destination traffic intensities from link data. *Journal of the American Statistical Association*, 91(433):365–377, 1996. doi:[10.1080/01621459.1996.10476697](https://doi.org/10.1080/01621459.1996.10476697).
- [28] A. Chen, J. Cao, and T. Bu. Network tomography: Identifiability and Fourier domain estimation. *IEEE Transactions on Signal Processing*, 58(12):6029–6039, 2010. doi:[10.1109/TSP.2010.2068294](https://doi.org/10.1109/TSP.2010.2068294).
- [29] H. Lev-Ari, Y. Ephraim, and B. L. Mark. Traffic rate network tomography with higher-order cumulants. *Networks*, 81(2):220–234, 2023. doi:[10.1002/net.22127](https://doi.org/10.1002/net.22127).
- [30] L. Hartung, N. Holden, and Y. Peres. Trace reconstruction with varying deletion probabilities. In *Proceedings of the Fifteenth Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*, pages 54–61, 2018. doi:[10.1137/1.9781611975062.6](https://doi.org/10.1137/1.9781611975062.6). Open preprint: [arXiv:1708.02216](https://arxiv.org/abs/1708.02216).
- [31] S. Narayanan. Improved algorithms for population recovery from the deletion channel. In *Proceedings of the 2021 ACM–SIAM Symposium on Discrete Algorithms (SODA)*, pages 1259–1278, 2021. doi:[10.1137/1.9781611976465.77](https://doi.org/10.1137/1.9781611976465.77).
- [32] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski. Phase retrieval via matrix completion. *SIAM Journal on Imaging Sciences*, 6(1):199–225, 2013. doi:[10.1137/110848074](https://doi.org/10.1137/110848074).
- [33] P. McCullagh. *Tensor Methods in Statistics*. Dover Publications, 2018. [Author’s online edition](#).