A Rapid Estimation of Distributional Statistics in Probabilistic Data Structures and Cryptography

Alice K. Ng[1,2], Jiahua Xu[1,2], Paolo Tasca[2]

[1] University College London, Gower Street, London, WC1E 6BT, UK
[2] Exponential Science, Willow House, Cricket Square, Grand Cayman KY1-1001, Cayman Islands

Abstract. We develop a unified distributional framework for coverage and overlap in random set systems with fixed-cardinality, without-replacement sampling. On the exact side, we introduce a threshold-state dynamic program that tracks level counts (the numbers of items covered by exactly ℓ parties) and yields finite-N distributions for exactly-ℓ and at-least-t parties coverage. By aggregation, the same engine recovers the classical recursions for the union, the intersection, and their joint law. On the asymptotic side, we prove a multivariate Central Limit Theorem (CLT) for the full level-count vector via Stein's method with an exchangeable-pairs coupling; linear projections (including the union, the intersection, and their difference) are thus Gaussian in the limit, and smooth indices such as the Jaccard similarity admit delta-method approximations with closed-form moments. The exact evaluator supports moderate instances and provides ground truth for verification, while the Gaussian surrogates are accurate across broad regimes. In terms of cost, exact evaluation scales at worst between $O(mN^2)$ and $O(mN^3)$ depending on the target, whereas assembling the CLT requires only $O(m^3)$. We illustrate how these tools inform the design of probabilistic data structures (Bloom filters, MinHash), incidence-graph coverage planning, and cryptographic mechanisms such as threshold recovery and committee overlap.

Keywords: Random Sets · Distribution Statistics · Data Sets · Probabilistic Data Structures

1 Introduction

Random set systems are a basic combinatorial model with appearances in probabilistic data structures (e.g., Bloom filters and MinHash), incidence/coverage processes (e.g., bipartite graphs and sensor coverage), and cryptography (e.g., additive/threshold secret sharing over index sets). Consider m subsets $P_1, \ldots, P_m$ drawn from a finite ground set $[N] = \{1, \ldots, N\}$, where each P_r is chosen uniformly at random from $\binom{[N]}{n_r}$, independently across r (fixed–cardinality, without–replacement sampling). In many applications one needs not only expecta-

tions but the full distributions of

$$U = \Big|\bigcup_{r=1}^{m} P_r\Big| \quad \text{(total union)}, \qquad I = \Big|\bigcap_{r=1}^{m} P_r\Big| \quad \text{(total intersect)}, \tag{1}$$

as well as functionals such as the Jaccard similarity $J = I/U$.

A convenient way to organise these statistics is via the *level counts*

$$Z_\ell = |\{i \in [N] : i \text{ belongs to exactly } \ell \text{ of } P_1, \ldots, P_m\}|, \qquad \ell = 0, 1, \ldots, m.$$

Then many quantities of interest are linear projections of the level–count vector $\mathbf{Z} = (Z_0, \ldots, Z_m)$: e.g., $U = \sum_{\ell \geq 1} Z_\ell$, $I = Z_m$, $U - I = \sum_{\ell=1}^{m-1} Z_\ell$, and threshold counts $W_t = \sum_{\ell \geq t} Z_\ell$. This Z–view unifies (U, I) and a broad family of threshold or aggregate counts, and it naturally supports Gaussian approximations via multivariate limit theorems.

We study the fixed–size model with arbitrary $\{n_r\}$ and develop both exact and asymptotic tools around the same state representation.

Contributions.

- **Exact counting via a threshold state (section 4).** We introduce a Markovian threshold state that tracks bucket sizes by coverage level up to a threshold τ, yielding a forward dynamic program with matrix update $\mathbf{S}_{k+1}^{\langle\tau\rangle} = \mathbf{S}_k^{\langle\tau\rangle} + M_\tau\, \mathbf{n}_{k+1}^{\langle\tau\rangle}$, and multivariate–hypergeometric weights. This produces exact finite–N probability mass functions (PMFs) for threshold coverage W_t (take $\tau = t$) and level counts Z_ℓ (take $\tau = \ell{+}1$). We give tight one–step feasibility bounds for dynamic programming (DP) and show that, after aggregation, this engine *reduces* to the classical global recursions U, I, in Appendix A. These identities let us verify asymptotics directly against exact laws in section 4.
- **General multivariate CLT for level counts (section 5).** We prove a multivariate CLT for $(Z_0, \ldots, Z_m)$ with $O(N^{-1/2})$ accuracy. Any fixed linear projection $(u^\top Z, v^\top Z, \ldots)$, including (U, I), $U - I$, W_t, or multi-slice vectors, is thus asymptotically Gaussian with explicit means and covariances obtained by block–summing the covariance of Z.
- **Smooth indices via the delta method (section 6).** We give a delta–method Gaussian approximation (closed-form mean/variance) for $J = I/U$ and, more generally, for any smooth functional g of a linear projection of Z. We also provide the exact finite–N distribution of J, for any $m \geq 2$, via recursion function F_m.
- **Algorithms, validations, and use cases (section 7).** We implement a memoised evaluator for the threshold–state DP to calculate the exact PMFs, and compare it with the CLT/delta approximations across regimes, and illustrate design implications for Bloom filters under merges, MinHash planning, threshold secret-sharing, and incidence-graph coverage targets.

2 Related work

Unions, intersections, and fixed–cardinality models. Classical results give *univariate* laws and moments for unions (and, less often, intersections) under uniform without–replacement sampling, typically with equal subset sizes; see Barot–de la Peña [1] for unions and Kalinka [7] for intersections. Joint laws for union *and* intersection are not standard, and explicit treatments that allow arbitrary $\{n_r\}$ are scarce. Our work supplies a finite–N recursion for the joint law that handles general $\{n_r\}$ and provides sharp feasibility bounds; the univariate laws follow by marginalization.

Bernoulli presence models and Jaccard. A substantial line of work analyzes Jaccard under i.i.d. Bernoulli presence/absence, yielding exact nulls and asymptotics [5,8]. In the fixed–cardinality (hypergeometric) design relevant here, most exact results focus on m=2 (e.g., Real–Vargas [9]). We extend to arbitrary m via the joint recursion and, in a unified *level-count* view, provide Gaussian/delta approximations for Jaccard and other smooth indices.

Stein's method for multivariate normal approximation. Exchangeable-pairs approaches for vectors were developed by Chatterjee–Meckes [4] and Reinert–Röllin [10]. We instantiate these techniques in the fixed–cardinality set model: a swap coupling yields a bivariate CLT for (union, intersection) and a multivariate CLT for the full level-count vector, both with $O(N^{-1/2})$ rates, from which linear projections inherit Gaussian limits.

Probabilistic data structures and incidence graphs. Bloom filters and MinHash drive distributional questions about unions, intersections, and ratios; prior analyses often rely on independence or Poissonization heuristics. Our exact recursion and level-count CLTs offer finite-N design rules for merged filters and similarity sketches under without–replacement sampling. Related intersection/coverage viewpoints in random intersection graphs typically target graph-level asymptotics rather than exact finite-N joint laws for prescribed set sizes; our counts and CLTs can be repurposed for such incidence-graph design tasks.

Relative to (i) univariate union/intersection results with equal sizes, (ii) Bernoulli-based Jaccard analyses, and (iii) graph-level asymptotics, we provide: a computable finite-N joint recursion for union and intersection with arbitrary $\{n_r\}$; a unifying level-count framework with multivariate CLTs (and delta-method indices); and a threshold-state dynamic program whose aggregation recovers the classical univariate and bivariate recursions used for verification.

3 Model and notation

Let $[N] = \{1, 2, \ldots, N\}$ be a finite universe and let $m \in \mathbb{N}$ be fixed. For $r \in [m]$, party r samples a subset

$$P_r \subseteq [N], \qquad |P_r| = n_r,$$

uniformly without replacement, independently across r. We allow $n_r = n_r(N)$ to depend on N, and write

$$\alpha_r := \frac{n_r}{N} \in (0,1), \qquad r \in [m],$$

with α_r fixed as $N \to \infty$ (the "fixed–proportions" regime). For brevity, write $S_m = (n_1, \ldots, n_m)$ and $\alpha = (\alpha_1, \ldots, \alpha_m)$.

Level–count vector and linear projections. For each item $i \in [N]$, let the per-item coverage count be

$$R(i) \;:=\; \sum_{r=1}^{m} \mathbf{1}\{i \in P_r\}.$$

For $\ell = 0, 1, \ldots, m$ define the level counts

$$Z_\ell = |\{\, i \in [N] : R(i) = \ell\}|,$$

so $\mathbf{Z} = (Z_0, \ldots, Z_m)$ and $\sum_{\ell=0}^{m} Z_\ell = N$.

Many statistics of interest are linear projections of Z:

$$U = \sum_{\ell \geq 1} Z_\ell, \qquad I = Z_m, \qquad W_t := \sum_{\ell \geq t} Z_\ell,$$

and including the non–unanimous mass $U - I = \sum_{\ell=1}^{m-1} Z_\ell$, $A_P = \sum_{\ell \in P} Z_\ell$ $(P \subseteq \{0, \ldots, m\})$.

4 Exact counting via a threshold state

We study the exact laws of the level–count statistics

$$Z_\ell := \big|\{i : R_m(i) = \ell\}\big|, \qquad W_t := \big|\{i : R_m(i) \geq t\}\big| = \sum_{\ell \geq t} Z_\ell,$$

where for $k \in [m]$ the running coverage count of item i is

$$R_k(i) = \sum_{r=1}^{k} \mathbf{1}\{i \in P_r\}, \qquad R_0(i) := 0.$$

We process parties sequentially and keep a *threshold state* that makes the evolution Markov, so the next party's draw decomposes into independent multivariate–hypergeometric choices across buckets.

Threshold state. Fix an integer $\tau \in [1, m]$. After k parties, define

$$\mathbf{B}_k^{\langle\tau\rangle} = (b_k^{(0)}, \ldots, b_k^{(\tau-1)})^\top \in \mathbb{Z}_{\geq 0}^{\tau\times 1}, \quad \sum_{s=0}^{\tau-1} b_k^{(s)} \leq N,$$

with

$$b_k^{(s)} = \big|\{i : R_k(i) = s\}\big| \ (0 \leq s \leq \tau{-}1), \qquad b_k^{(\tau)} = \big|\{i : R_k(i) \geq \tau\}\big| = N{-}\sum_{s=0}^{\tau-1} b_k^{(s)}.$$

Thus the bucket $b_k^{(\tau)}$ aggregates all items at level τ or higher. This single parameter τ lets us read off different targets:

$$W_t = b_m^{(\tau)} \text{ when } \tau = t; \qquad Z_\ell = b_m^{(\ell)} \text{ when } \tau = \ell + 1.$$

Counts of ordered selections. Let $H_k(\mathbf{B}_k^{\langle\tau\rangle})$ denote the number of ordered k-tuples $(P_1, \ldots, P_k)$ that realise the state $\mathbf{B}_k^{\langle\tau\rangle}$. The base case is deterministic:

$$\mathbf{B}_1^{\langle\tau\rangle} = (N - n_1, n_1, 0, \ldots, 0)^\top, \qquad H_1(\mathbf{B}_1^{\langle\tau\rangle}) = \binom{N}{n_1},$$

and $H_1(\cdot) = 0$ for any other state.

One–step update. Given $\mathbf{B}_k^{\langle\tau\rangle}$ and adding party $k{+}1$ with $|P_{k+1}| = n_{k+1}$, split that party's draw by buckets as

$$\mathbf{n}_{k+1}^{\langle\tau\rangle} = (n_{k+1}^{(0)}, \ldots, n_{k+1}^{(\tau-1)})^\top \in \mathbb{Z}_{\geq 0}^{\tau\times 1}, \quad n_{k+1}^{(\tau)} := n_{k+1}{-}\sum_{s=0}^{\tau-1} n_{k+1}^{(s)}, \quad 0 \leq n_{k+1}^{(s)} \leq b_k^{(s)} \quad \forall s \in [0, \tau].$$

Elements drawn from bucket $s < \tau$ move up exactly one level; elements drawn from bucket τ stay in bucket τ. Hence

$$b_{k+1}^{(0)} = b_k^{(0)} - n_{k+1}^{(0)}, \tag{2a}$$
$$b_{k+1}^{(s)} = b_k^{(s)} - n_{k+1}^{(s)} + n_{k+1}^{(s-1)} \quad (1 \leq s \leq \tau - 1), \tag{2b}$$
$$b_{k+1}^{(\tau)} = b_k^{(\tau)} + n_{k+1}^{(\tau-1)}. \tag{2c}$$

We can write

$$\mathbf{B}_{k+1}^{\langle\tau\rangle} = \mathbf{B}_k^{\langle\tau\rangle} + M_\tau \mathbf{n}_{k+1}^{\langle\tau\rangle} \tag{3}$$

where $M_\tau \in \mathbb{Z}^{\tau\times\tau}$, $M_{j,j} = -1$ for $j \in [1, \tau]$, and $M_{j,j-1} = +1$ for $j \in [2, \tau]$, i.e.

$$M_\tau = \begin{pmatrix} -1 & 0 & \cdots & 0 \\ 1 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & 0 \\ \vdots & \cdots & 1 & -1 \end{pmatrix} \in \mathbb{Z}^{\tau\times\tau}. \tag{4}$$

Since M_τ is lower triangular with diagonal -1, it is invertible and the draw split is uniquely recovered from the state difference:

$$\mathbf{n}_{k+1}^{\langle\tau\rangle} = M_\tau^{-1}\left(\mathbf{B}_{k+1}^{\langle\tau\rangle} - \mathbf{B}_k^{\langle\tau\rangle}\right). \tag{5}$$

Componentwise, with $\Delta b^{(s)} := b_{k+1}^{(s)} - b_k^{(s)}$,

$$\Delta b^{(0)} = -n_{k+1}^{(0)}, \qquad \Delta b^{(s)} = -n_{k+1}^{(s)} + n_{k+1}^{(s-1)} \ (1 \le s \le t-1).$$

Since M_τ^{-1} is the lower triangle filled with -1,

$$n_{k+1}^{(s)} = -\sum_{r=0}^{s} \Delta b^{(r)} \quad (0 \le s \le t-1), \qquad n_{k+1}^{(t)} = n_{k+1} - \sum_{s=0}^{t-1} n_{k+1}^{(s)}.$$

Given a feasible split, the number of ways to realise it is the multivariate–hypergeometric factor

$$\prod_{s=0}^{\tau} \binom{b_k^{(s)}}{n_{k+1}^{(s)}}.$$

Therefore

$$H_{k+1}(\mathbf{B}_{k+1}^{\langle\tau\rangle}) = \sum_{\mathbf{B}_k^{\langle\tau\rangle}} \left(\prod_{s=0}^{\tau} \binom{b_k^{(s)}}{n_{k+1}^{(s)}} \right) H_k(\mathbf{B}_k^{\langle\tau\rangle}), \tag{6}$$

summing over all the feasible predecessors $\mathbf{B}_k^{\langle\tau\rangle}$ consistent with boundaries (7a-e).

Feasible bounds. A transition $\mathbf{B}_k^{\langle\tau\rangle} \to \mathbf{B}_{k+1}^{\langle\tau\rangle}$ is feasible iff

$$0 \le n_{k+1}^{(s)} \le b_k^{(s)} \ (0 \le s \le \tau), \tag{7a}$$

$$\max\{0, b_k^{(0)} - n_{k+1}\} \ \le \ b_{k+1}^{(0)} \ \le \ b_k^{(0)}, \tag{7b}$$

$$\max\{0, b_k^{(s)} - n_{k+1}\} \ \le \ b_{k+1}^{(s)} \ \le \ b_k^{(s)} + \min\{b_k^{(s-1)}, n_{k+1}\} \quad (1 \le s \le \tau - 1), \tag{7c}$$

$$b_k^{(\tau)} \ \le \ b_{k+1}^{(\tau)} \ \le \ b_k^{(\tau)} + \min\{b_k^{(\tau-1)}, n_{k+1}\}, \tag{7d}$$

$$\sum_{r=0}^{s} b_k^{(r)} - \min\{b_k^{(s)}, n_{k+1}\} \ \le \ \sum_{r=0}^{s} b_{k+1}^{(r)} \le \ \sum_{r=0}^{s} b_k^{(r)} \qquad (0 \le s \le \tau - 1). \tag{7e}$$

Theorem 1 (PMF for threshold coverage W_t). *Fix $t \ge 1$ and take $\tau = t$. Then $W_t = b_m^{(\tau)}$, and*

$$\mathbb{P}(W_t = w) = \sum_{\mathbf{B}_m^{\langle t\rangle}:\ b_m^{(t)} = w} \frac{H_m(\mathbf{B}_m^{\langle t\rangle})}{\prod_{r=1}^{m} \binom{N}{n_r}}.$$

Theorem 2 (PMF for the level count Z_ℓ). *Fix $\ell \geq 0$ and take $\tau = \ell + 1$. Then $Z_\ell = b_m^{(\ell)}$, and*

$$\mathbb{P}(Z_\ell = z) = \sum_{\mathbf{B}_m^{\langle \ell+1 \rangle}:\ b_m^{(\ell)} = z} \frac{H_m(\mathbf{B}_m^{\langle \ell+1 \rangle})}{\prod_{r=1}^m \binom{N}{n_r}}.$$

Remark. For fixed τ, complexity is pseudo-polynomial in the size of the feasible grid of $\mathbf{B}_k^{\langle \tau \rangle}$ and practical for small m, N. Aggregating the state yields 1D/2D recursions for specific targets (below), with worst-case time $T(m, N)$ between $O(mN^2)$ and $O(mN^3)$.

Lemma 1 (Reduction to the union ($U = \sum_{l \geq 1}^m Z_l$)). *Set $\tau = 1$ so $\mathbf{B}_k^{\langle 1 \rangle} = (b_k^{(0)})$ and $U_k := N - b_k^{(0)} = |\cup_{r \leq k} P_r|$. Define aggregated counts from the threshold state*

$$C_k(u_k) := \sum_{\mathbf{B}_k^{\langle 1 \rangle}:\, N - b_k^{(0)} = u_k} H_k(\mathbf{B}_k^{\langle 1 \rangle}).$$

Then $C_1(u_1) = \binom{N}{n_1} \mathbf{1}\{u_1 = n_1\}$ and, for $k \geq 1$,

$$C_{k+1}(u_{k+1}) = \sum_{u_k} \binom{u_k}{n_{k+1} + u_k - u_{k+1}} \binom{N - u_k}{u_{k+1} - u_k} C_k(u),$$

Hence $\mathbb{P}(U_m = u_m) = C_m(u_m) / \prod_{r=1}^m \binom{N}{n_r}$.

Proof sketch. *With $\tau = 1$ the split is $(n^{(0)}, n^{(1)})$ between "new" (from $N - U_k$) and "already covered" (from U_k), and $U_{k+1} = U_k + n^{(0)}$ forces $n^{(0)} = u_{k+1} - u_k$. The multiplicity of a step is $\binom{U_k}{n_{k+1} - n^{(0)}} \binom{N - U_k}{n^{(0)}} = \binom{u_k}{n_{k+1} + u_k - u_{k+1}} \binom{N - u_k}{u_{k+1} - u_k}$, summing over feasible x gives the stated recursion. Detailed bounds for u_k and the full proof are in Appendix A.1. Note that the time complexity for the calculation is $T(m, N) = O(mN^2)$.* □

Lemma 2 (Reduction to the intersection ($I = Z_m$)). *Run the threshold scheme with the time–varying choice $\tau = k$ at step k, so $I_k := b_k^{(k)} = |\cap_{r \leq k} P_r|$. Define aggregated counts*

$$D_k(v_k) := \sum_{\mathbf{B}_k^{\langle k \rangle}:\, b_k^{(k)} = v_k} H_k(\mathbf{B}_k^{\langle k \rangle}).$$

Then $D_1(y; S_1) = \binom{N}{n_1} \mathbf{1}\{y = n_1\}$ and, for $k \geq 1$,

$$D_{k+1}(v_{k+1}) = \sum_{v_k} \binom{v_k}{v_{k+1}} \binom{N - v_k}{n_{k+1} - v_{k+1}} D_k(v_k).$$

Hence $\mathbb{P}(I_m = v_m) = D_m(v_m) / \prod_{r=1}^m \binom{N}{n_r}$.

Proof. *At step k, items with coverage $\geq k$ are exactly the current intersection, of size v_k. To have $I_{k+1} = v_{k+1}$, party $k+1$ must pick exactly v_{k+1} from those v_k*

and the remaining $n_{k+1} - v_{k+1}$ from outside, giving $\binom{v_k}{v_{k+1}}\binom{N-v_k}{n_{k+1}-v_{k+1}}$, summing over feasible v_k yields the recursion. Detailed bounds for u_k and the full proof are in Appendix A.2. Same as union, the time complexity for the calculation is $T(m, N) = O(mN^2)$. □

Lemma 3 (Reduction to the bivariate recursion F_m). *Fix $\tau = m$ and define aggregated counts from the threshold state by*

$$F_k(u_k, v_k) := \sum_{\mathbf{B}_k^{\langle m\rangle}:\ N-b_k^{(0)}=u_k,\ \sum_{s=k}^{m-1} b_k^{(s)}+b_k^{(\geq m)}=v_k} H_k(\mathbf{B}_k^{\langle m\rangle}).$$

Then $F_1(u_1, v_1) = \binom{N}{n_1}\mathbf{1}\{u_1 = v_1 = n_1\}$, and for all $m \geq 2$,

$$F_m(u_m, v_m) = \sum_{v_{m-1}} \sum_{u_{m-1}} \binom{v_{m-1}}{v_m} \binom{u_{m-1} - v_{m-1}}{n_m + u_{m-1} - u_m - v_m} \binom{N - u_{m-1}}{u_m - u_{m-1}} F_{m-1}(u_{m-1}, v_{m-1}),$$

Hence $p_{X,Y}(u_m, v_m) = F_m(u_m, v_m; S_m) \big/ \prod_{r=1}^m \binom{N}{n_r}$ as in (25).

The feasibility bounds on (u_{m-1}, v_{m-1}) and the full proof are in Appendix A.3. Note that the time complexity for the calculation is $T(m, N) = O(mN^3)$.

Proposition 1 (Jaccard index using bivariate recursion F_m). *Let $a \geq 1$, $0 \leq b \leq a$ be coprime. Then for any $m \geq 2$*

$$\mathbb{P}\left(J = \frac{b}{a}\right) = \sum_{k \in \mathcal{K}(a,b)} \frac{F_m(ka,\, kb)}{\prod_{r=1}^m \binom{N}{n_r}},$$

where $\mathcal{K}(a, b) = \{k \in \mathbb{Z}_{\geq 0} :\ (ka, kb)$ is feasible$\}$.

5 General multivariate CLT

This section provides the Gaussian approximation for the level–count vector $\mathbf{Z} = (Z_0, \ldots, Z_m)$. All normal approximations for univariate and multivariate projections (e.g. U, I, (U, I), W_t, or any (A, B)) then follow as corollaries by block–summing the covariance.

5.1 Means and one-item probabilities

For a single item i, independence across parties gives, for $\ell = 0, \ldots, m$,

$$p_\ell(\alpha) \ :=\ \mathbb{P}\{R(i) = \ell\} = [t^\ell] \prod_{r=1}^m \big((1 - \alpha_r) + \alpha_r t\big). \tag{8}$$

Here $[t^\ell]$ means "coefficient of t^ℓ" in a probability generating function (PGF). Identity (8) is the *coefficient-of* form of the elementary symmetric polynomials e_ℓ:

$$\mathbb{E}(t^{R(i)}) = \prod_{r=1}^m \big((1 - \alpha_r) + \alpha_r t\big) = \sum_{\ell=0}^m \Big(\prod_{r=1}^m (1 - \alpha_r)\Big) e_\ell\Big(\tfrac{\alpha_1}{1-\alpha_1}, \ldots, \tfrac{\alpha_m}{1-\alpha_m}\Big) t^\ell.$$

Hence p_ℓ is explicit and depends on α only, and computing all p_ℓ via a rolling 2-term convolution takes $T(m) = O(m^2)$.

By exchangeability of items,

$$\mathbb{E}[Z_\ell] \;=\; \sum_{i=1}^{N} \mathbb{P}\{R(i) = \ell\} \;=\; N\, p_\ell(\alpha). \tag{9}$$

5.2 Finite-N covariances: exact identities

Let $I_i^{(\ell)} \;:=\; \mathbf{1}\{R(i) = \ell\}$, so $Z_\ell = \sum_{i=1}^{N} I_i^{(\ell)}$. For distinct items $u \neq v$, write

$$q_{a,b} \;:= \mathbb{P}\big(R(u) = a,\; R(v) = b\big), \qquad a, b \in \{0, \ldots, m\}.$$

Then by expanding variances and covariances of sums of indicators:

$$\mathrm{Var}(Z_a) = \sum_{i=1}^{N} \mathrm{Var}(I_i^{(a)}) + \sum_{u \neq v} \mathrm{Cov}(I_u^{(a)}, I_v^{(a)}) = N\, p_a(1 - p_a) \;+\; N(N-1)\,\big(q_{a,a} - p_a^2\big), \tag{10}$$

$$\begin{aligned} \mathrm{Cov}(Z_a, Z_b) &= \sum_{u \neq v} \mathrm{Cov}(I_u^{(a)}, I_v^{(b)}) \;-\; N\, p_a p_b \\ &= N(N-1)\, q_{a,b} \;-\; N^2 p_a p_b, \qquad a \neq b. \end{aligned} \tag{11}$$

For a *fixed* party r, the pair $(\mathbf{1}\{u \in P_r\}, \mathbf{1}\{v \in P_r\})$ has the *hypergeometric* probabilities

$$\pi_{00}^{(r)} = \frac{(N - n_r)(N - n_r - 1)}{N(N-1)}, \quad \pi_{10}^{(r)} = \pi_{01}^{(r)} = \frac{n_r(N - n_r)}{N(N-1)}, \quad \pi_{11}^{(r)} = \frac{n_r(n_r - 1)}{N(N-1)}.$$

Define the per-party bivariate PGF

$$\phi_r(z, w) \;:=\; \pi_{00}^{(r)} + \pi_{10}^{(r)} z + \pi_{01}^{(r)} w + \pi_{11}^{(r)} zw.$$

Independence across parties yields

$$q_{a,b} \;=\; [z^a w^b] \prod_{r=1}^{m} \phi_r(z, w). \tag{12}$$

Cost. All $q_{a,b}$ can be obtained by m successive 2×2 convolutions on an $(\leq m{+}1) \times (\leq m{+}1)$ table, with time complexity $T(m) = O(m^3)$. Plugging into (10)–(11) forms $\Sigma_N = \mathrm{Cov}(\mathbf{Z})$ in the same $T(m) = O(m^3)$.

5.3 Asymptotic covariance scaling

Expanding each ϕ_r to first order in $(N-1)^{-1}$ gives

$$\phi_r(z, w) = \big((1 - \alpha_r) + \alpha_r z\big)\big((1 - \alpha_r) + \alpha_r w\big) + \frac{\alpha_r(1 - \alpha_r)}{N - 1}\,\big(-1 + z + w - zw\big).$$

Multiplying over r and extracting coefficients as in (12) yields

$$q_{a,b} = \; p_a p_b \; + \; \frac{1}{N-1}\, S_{a,b}(\alpha) \; + \; O\big((N-1)^{-2}\big),$$

where $S_{a,b}(\alpha)$ is an $O(1)$ quantity depending only on α. Substituting into (10)–(11) gives

$$\Sigma_N \;=\; N\,\Sigma(\alpha) \;+\; O(1), \tag{13}$$

with entries

$$\Sigma_{aa}(\alpha) = p_a(1-p_a) \;-\; \sum_{r=1}^{m} \alpha_r(1-\alpha_r)\,\big(\Delta_a^{(-r)}\big)^2, \tag{14}$$

$$\Sigma_{ab}(\alpha) = -p_a p_b \;-\; \sum_{r=1}^{m} \alpha_r(1-\alpha_r)\,\Delta_a^{(-r)}\,\Delta_b^{(-r)}, \qquad a \neq b. \tag{15}$$

Here $p_\ell^{(-r)} := [t^\ell] \prod_{s\neq r}\big((1-\alpha_s)+\alpha_s t\big)$, $\Delta_\ell^{(-r)} := p_\ell^{(-r)} - p_{\ell-1}^{(-r)}$, and $p_{-1}^{(-r)} := 0$. All $p_k^{(-r)}$ for $r = 1..m$, $\ell = 0..m$ can be computed in $O(m^2)$ time; assembling $\Sigma(\alpha)$ via (14)–(15) requires $O(m^3)$ time (due to $O(m^2)$ pairs (a,b) each with an $O(m)$ inner sum).

5.4 A multivariate CLT for the level counts

Because $\sum_{l=0}^{m} Z_\ell = N$ is deterministic, Σ_N is singular in the direction of the all-ones vector $\mathbf{1} = (1,\ldots,1)^\top$. Let $\mathcal{T} := \{x \in \mathbb{R}^{m+1} : \; \mathbf{1}^\top x = 0\}$ be the m-dimensional *tangent subspace* and let Π be the orthogonal projector onto $\mathcal{T}$. Define the centered, scaled vector

$$W_N \;:=\; \frac{1}{\sqrt{N}}\,(\Pi\,\Sigma(\alpha)\,\Pi^\top)^{-1/2}\,\Pi\,\big(Z - \mathbb{E}(Z)\big).$$

Theorem 3 (Multivariate CLT on the simplex tangent). *Fix m and $\alpha_r \in (0,1)$. Then $W_N \to \mathcal{N}(0, I_m)$ as $N \to \infty$.*
$\lim_{N\to\infty} W_N \sim \mathcal{N}(0, I_m)$.
Moreover, for smooth test functions the approximation error is $O(N^{-1/2})$.

Proof (sketch). We build an exchangeable pair $(\mathbf{Z}, \mathbf{Z}')$ by a swap that only modifies $O(1)$ items while preserving all set sizes $|P_r| = n_r$: pick a random item K; for each party r, if $K \in P_r$ swap it with a uniformly chosen element outside P_r, and if $K \notin P_r$ swap it with a uniformly chosen element inside P_r. This creates small increments $\Delta := \mathbf{Z}' - \mathbf{Z}$ with bounded third moments and gives the regression

$$\mathbb{E}(\Delta \mid \mathbf{Z}) \;=\; -\frac{1}{N}\Pi\,(\mathbf{Z} - \mathbb{E}(\mathbf{Z})),$$

where Π is the projector onto T. Conditional second moments match $\Sigma(\alpha)$ up to $O(1/N)$. The multivariate Stein–exchangeable-pairs theorem then yields the CLT with $O(N^{-1/2})$ error. Full details in Appendix C.

6 Linear projections and smooth indices

This section records the principal projections of the level counts $\mathbf{Z} = (Z_0, \dots, Z_m)$ and their Gaussian/delta approximations. All limits are immediate from Theorem 3. Exact finite–N pmfs come from the recursions in Appendix A; figures are in Appendix D.

Corollary 1 (CLTs for linear projections). *Let L be any fixed $\ell \times (m+1)$ matrix and set $X^{(\ell)} := L\mathbf{Z}$. Then $X^{(\ell)}$ inherits a joint Gaussian limit. Under Theorem 3,*

$$\frac{X^{(\ell)} - \mathbb{E}(X)^{(\ell)}}{\sqrt{N}} \to \mathcal{N}_\ell\big(0,\ L\Sigma(\alpha)L^\top\big).$$

In particular, $U = \sum_{k\geq 1} Z_k$ and $I = Z_m$ are asymptotically normal, and (U, I) is asymptotically bivariate normal with covariance $L\Sigma(\alpha)L^\top$ for the selector L that picks $(\sum_{k\geq 1} Z_k,\ Z_m)$.

Once $\Sigma(\alpha)$ is precomputed in $O(m^3)$ time (Section 5.3), forming $L\Sigma(\alpha)L^\top$ costs $T(m) = O(\ell m^2)$.

Lemma 4 (Delta–method for smooth indices). *Let $g : \mathbb{R}^\ell \to \mathbb{R}$ be C^2 near $\mathbb{E}(X)^{(\ell)}$ and let $T = g\big(X^{(\ell)}\big)$. Then*

$$\mathbb{E}(T) = g(\mu) + \tfrac{1}{2}\operatorname{tr}\big(H_g(\mu)\,\Sigma_L\big) + O(N^{-1/2}), \quad \mathit{Var}(T) = \nabla g(\mu)^\top \Sigma_L \nabla g(\mu) + O(N^{-1/2}),$$

where $\mu = \mathbb{E}(X)^{(\ell)}$ and $\Sigma_L = L\Sigma(\alpha)L^\top$, and $(T - \mathbb{E}(T))/\sqrt{\mathit{Var}(T)} \Rightarrow \mathcal{N}(0,1)$.

*Cost.*Given Σ_L, evaluating the mean/variance expansions above is $O(\ell^2)$

6.1 Example: Univariate U and I

Means. By (9) and the identities in §5.1,

$$\mu_U := \mathbb{E}(U) = N\Big(1 - \prod_{r=1}^{m}(1-\alpha_r)\Big), \qquad \mu_I := \mathbb{E}(I) = N\prod_{r=1}^{m}\alpha_r.$$

Variances (from $\Sigma(\alpha)$). Let $L_U \in \mathbb{R}^{1\times(m+1)}$ be $L_U = (0, 1, \dots, 1)$ and $L_I = (0, ..., 0, 1)$. Then, from (13)–(15),

$$\operatorname{Var}(U) = N\,v_U(\alpha) + O(1), \quad v_U(\alpha) = L_U\,\Sigma(\alpha)\,L_U^\top, \tag{16}$$

$$\operatorname{Var}(I) = N\,v_I(\alpha) + O(1), \quad v_I(\alpha) = L_I\,\Sigma(\alpha)\,L_I^\top. \tag{17}$$

Corollary 2 (Univariate CLTs as projections of Theorem 3). *With $U = \sum_{\ell\geq 1} Z_\ell$ and $I = Z_m$,*

$$\frac{U - \mu_U}{\sqrt{N}} \to \mathcal{N}\big(0,\ v_U(\alpha)\big), \qquad \frac{I - \mu_I}{\sqrt{N}} \to \mathcal{N}\big(0,\ v_I(\alpha)\big),$$

with an $O(N^{-1/2})$ error for smooth test functions.

*Cost.*After $\Sigma(\alpha)$, obtaining v_U, v_I is $O(m^2)$; computing μ_U, μ_I is $O(m)$.

6.2 Example: Bivariate (U, I)

Mean vector and covariance (as a projection). Let $L \in \mathbb{R}^{2\times(m+1)}$ have first row L_U and second row L_I. Then

$$E\begin{bmatrix}U\\I\end{bmatrix} = \begin{bmatrix}\mu_U\\\mu_I\end{bmatrix}, \qquad \mathrm{Cov}\begin{bmatrix}U\\I\end{bmatrix} = N\,\Sigma_{UI} + O(1), \qquad \Sigma_{UI} := L\,\Sigma(\alpha)\,L^\top.$$

(Explicit formulas follow by substituting (14)–(15).)

Corollary 3 (Bivariate CLT for (U, I)). *Under Theorem 3,*

$$\frac{1}{\sqrt{N}}\begin{pmatrix}U-\mu_U\\I-\mu_I\end{pmatrix} \to \mathcal{N}_2\big(0,\ \Sigma_{UI}\big),$$

with an $O(N^{-1/2})$ error for smooth test functions (and hence in convex–set distance).

Cost. Forming Σ_{UI} is $O(m^2)$ once $\Sigma(\alpha)$ is available.

6.3 Example: Jaccard indices from (U, I)

The Jaccard index $J = U/I \in [0, 1]$ is a smooth function of (U, I) on the set $\{u > 0\}$. In our model $\mu_U > 0$, so the delta method applies.

Corollary 4 (Delta–method approximation for J). *Let $g(u, v) = v/u$ and write $\mu = (\mu_U, \mu_I)$, Σ_{UI} as above. Then*

$$\frac{J-\mu_J}{\sigma_J} \ \Rightarrow\ \mathcal{N}(0,1), \qquad \textit{with error } O(N^{-1/2})$$

for smooth test functions, where

$$\mu_J = g(\mu) \ +\ \tfrac{1}{2}\,\mathrm{tr}\big(H_g(\mu)\,\Sigma_{UI}\big) \ +\ O(N^{-1/2}) = \frac{\mu_I}{\mu_U} \ +\ \frac{\mu_I\ Var(U) - \mu_U\ Cov(U,I)}{\mu_U^3} \ +\ O(N^{-1/2}), \tag{18}$$

$$\sigma_J^2 = \nabla g(\mu)^\top \Sigma_{UI}\,\nabla g(\mu) \ +\ O(N^{-1/2}) = \frac{Var(I)}{\mu_U^2} + \frac{\mu_I^2}{\mu_U^4}\,Var(U) - \frac{2\mu_I}{\mu_U^3}\,Cov(U,I) \ +\ O(N^{-1/2}). \tag{19}$$

Here $\nabla g(\mu) = (-\mu_I/\mu_U^2,\ 1/\mu_U)$ *and* $H_g(\mu) = \begin{bmatrix}2\mu_I/\mu_U^3 & -1/\mu_U^2\\ -1/\mu_U^2 & 0\end{bmatrix}$.

Proof (sketch). By Corollary 3, (U, I) is asymptotically normal with mean μ and covariance $N\,\Sigma_{UI} + O(1)$. Apply the multivariate delta method to $g(x = u, v) = v/u$, expanding mean and variance to first order in $1/N$ to obtain (18)–(19).

Cost. After Σ_{UI}, evaluating (18)–(19) is $O(1)$.

7 Applications

We show how the exact counts and the moment/CLT results for (U, I) and J translate into practical procedures. Means/variances/covariances come from section 6. When exact probabilities are feasible we use the dynamic programs of section 4; otherwise we use the (uni/bivariate) normal surrogates with standard continuity corrections.

7.1 Probabilistic data structures

probabilistic data structures (PDS) are compact, mergeable sketches (e.g., Bloom filters, MinHash). Our laws for (U, I) and J yield finite–N design rules.

Bloom filters: sizing under merges Consider m parties that insert $|P_r| = n_r$ distinct elements into a common Bloom filter [3] of length M with h hash functions (no deletions). The false-positive rate (FPR), conditional on the *union* size U, is well-approximated by

$$\mathrm{FPR}(U) \;\approx\; \left(1 - e^{-hU/M}\right)^h.$$

Hence, for a target FPR: ε and reliability $1-\delta$, choose (M, h) so that

$$\mathrm{P}\big(\mathrm{FPR}(X) \le \varepsilon\big) = \mathrm{P}\left(X \le \frac{M}{h} \log \frac{1}{1-\varepsilon^{1/h}}\right) \;\ge\; 1-\delta. \tag{20}$$

Two evaluation routes:

- *Exact:* compute the RHS by summing the counts for union C_m: $\sum_{u \le u^\star} \frac{C_m(u)}{\prod_r \binom{N}{n_r}}$, with $u^\star$ the threshold inside the braces. Complexity $T(m, N) = O(mN^2)$
- *Gaussian:* use $U \approx \mathcal{N}(\mu_U, \sigma_U^2)$ and $\mathrm{P}(U \le u^\star) \approx \Phi\big((u^\star + 0.5 - \mu_U)/\sigma_U\big)$. Complexity $T(m) = O(m^3)$ to precompute $\Sigma(\alpha)$.

For merge-heavy workloads, you can invert (20) to solve for the minimal M given h (or vice versa) at reliability $1-\delta$.

MinHash: sample size planning with a J prior A MinHash sketch [6] with T independent hash functions yields $\widehat{J} = \frac{1}{T}\sum_{t=1}^{T} B_t$ with $B_t \sim \mathrm{Bernoulli}(J)$ conditionally on J. Our exact/approximate laws for J supply a *prior* (or design distribution) for planning T.

- *Frequentist sizing:* to guarantee a margin η at confidence $1-\delta$ for a nominal J_0, take $T \;\ge\; \left(\frac{z_{1-\delta/2}}{\eta}\right)^2 J_0(1-J_0)$. Using $J_0 = \mu_J$ from (18) is a principled default.
- *Bayesian credible intervals:* treat the law of J (exact for $m = 2$, delta-normal otherwise) as a prior to get posterior bands for J given $\widehat{J}$; this absorbs the finite-N overlap structure among parties.

7.2 Secret sharing and access structures

In secret sharing protocols [2, 11], a dealer splits a secret D into N labelled *shares* and distributes them to parties. Operationally, multiple shares may be assigned to a single holder ("sharding"). In our model, party r receives n_r distinct shares uniformly without replacement, so different parties can receive overlapping shares.

In a threshold (k, N) scheme, recovery depends on the number of *unique* shares held by a group: even if the group holds more than k shares in total, overlaps can leave fewer than k distinct shares, making reconstruction impossible. Thus shares should be spread broadly so that any k independent parties can recover the secret; excessive overlap undermines both availability and security.

Threshold (k, N) schemes. Recovery occurs iff the union of allocated shares across the group reaches size k, i.e. $U = \left|\cup_r P_r\right| \geq k$. Evaluate

$$\mathbb{P}(U \geq k) = \sum_{u=k}^{N} \sum_{v} \frac{F_m(u, v)}{\prod_r \binom{N}{n_r}} \quad \text{or} \quad \mathbb{P}(U \geq k) \approx 1 - \Phi\Big(\frac{k - 0.5 - \mu_U}{\sigma_U}\Big).$$

To also control *over-concentration* (too many identical shares), impose $I \leq l$ simultaneously and compute $\mathbb{P}(U \geq k,\ I \leq l)$ either exactly by summing F_m on $\{u \geq k,\, v \leq l\}$, or via the bivariate normal CDF for (U, I).

All-or-nothing (N, N) additive sharing (special case). Here recovery requires full coverage, $U = N$. Compute

$$\mathbb{P}(U = N) = \sum_{v} \frac{F_m(N, v)}{\prod_r \binom{N}{n_r}} \quad \text{or} \quad \mathbb{P}(U = N) \approx 1 - \Phi\Big(\frac{N - 0.5 - \mu_U}{\sigma_U}\Big).$$

7.3 Incidence-graph operations

View the system as a bipartite incidence graph with left vertices $[N]$ (items) and right vertices $[m]$ (parties); U counts covered items and I the fully redundant items. The bivariate law F_m and its Gaussian surrogate enable compact joint guarantees and inference.

Joint SLAs: cover enough, avoid hotspots Pick targets k and l and certify

$$\mathbb{P}(U \geq k,\ I \leq l) \ \geq\ 1 - \delta.$$

Evaluate either by summing F_m on the rectangle $\{u \geq k,\, v \leq l\}$ or with the bivariate normal CDF over $[k - 0.5, \infty) \times (-\infty, l + 0.5]$. You can invert this numerically to choose $\{n_r\}$ (or m) at reliability $1 - \delta$.

Conditional redundancy from observed coverage Given an observed coverage $\widehat{x}$ (e.g., via Bloom filter bit density), the bivariate normal gives

$$\mathbb{E}(I \mid U = \widehat{u}) = \mu_I + \rho \frac{\sigma_I}{\sigma_U}(\widehat{u} - \mu_U), \qquad \mathrm{Var}(I \mid U = \widehat{u}) = \sigma_I^2(1 - \rho^2),$$

with $\rho = \mathrm{Cov}(U, I)/(\sigma_U \sigma_I)$. Use $\mathrm{P}(I > l \mid U = \widehat{u})$ as an online alarm for redundancy spikes or collusion risk.

7.4 Blockchain and distributed consensus

Blockchain data availability and sharding. With erasure-coded blocks (need any k of N chunks), assign chunks to validators as P_r. Reconstructability is $\mathbb{P}(U \geq k)$; redundancy and sampling fairness are controlled by I and by tails of Z_ℓ/W_t. This provides explicit, finite-N guarantees for data-availability sampling and shard replication.

Blockchain protocols: PoW, PoS, PoB (selection overlap and propagation). When stake-weighted samplings occur over epochs, our intersection machinery (D_m) measures expected and tail overlap of validator sets across epochs (sybil/capture risk), and Z_ℓ tails quantify how many validators are repeatedly selected.

Distributed consensus protocols (committee overlap and quorum safety). Across rounds, committees are subsets; the safety of BFT-style protocols hinges on sufficient *intersection* of honest quorums. Model two (or multiple) committees as parties and evaluate the intersection distribution with D_m (using Lemma 2) or via the CLT for (U, I) when N is large. This quantifies the probability of insufficient overlap (safety) or excessive overlap (centralisation risk).

8 Conclusion

We developed a unified distributional framework for random set systems under fixed–cardinality, without–replacement sampling. The organising principle is the *level–count* vector $\mathbf{Z} = (Z_0, \ldots, Z_m)$, where Z_ℓ counts items covered by exactly ℓ parties. Classic statistics are linear views of Z, for example $U = \sum_{\ell \geq 1} Z_\ell$, $I = Z_m$, $W_t = \sum_{\ell \geq t} Z_\ell$.

We introduced a Markovian threshold state whose one–step update is linear, $\mathbf{S}_{k+1}^{\langle\tau\rangle} = \mathbf{S}_k^{\langle\tau\rangle} + M_\tau \, \mathbf{n}_{k+1}^{\langle\tau\rangle}$, with multivariate–hypergeometric weights and sharp feasibility bounds. This yields exact finite–N PMFs for W_t (take $\tau = t$) and Z_ℓ (take $\tau = \ell{+}1$). After aggregation, the same engine collapses to the global recursions e.g. $\tau = 1 \Rightarrow C_m$ (union), $\tau = k \Rightarrow D_m$ (intersection at step k), and $\tau = m \Rightarrow F_m$ (bivariate (U, I)).

Using an exchangeable–pairs Stein coupling we proved a multivariate CLT for Z with $O(N^{-1/2})$ accuracy, which immediately gives asymptotic normality for any fixed linear projection of Z (including (U, I), $U{-}I$, and W_t), and delta–method surrogates for smooth indices such as the Jaccard $J = U/I$.

Closed-form (finite-N) expressions for means and covariances come from one- and two-item probabilities.

The exact evaluators are pseudo-polynomial in N (feasible grid size), while the CLT depends only on m. Concretely, worst-case update counts are $T = O(mN^2)$ for univariate cases and $T = O(mN^3)$ for bivariate cases. By contrast, assembling the CLT needs only the first– and second–order structure of Z: the one-item pgf yields p_ℓ in $O(m^2)$, and the two-item pgf yields all $p^{(2)}_{\ell,\ell'}$ in $O(m^3)$; hence *full* $(m{+}1) \times (m{+}1)$ covariance construction is $O(m^3)$, independent of N. Thus for large N the workload drops from $T = O(mN^2)$ (or worse) to $T = O(m^3)$ when using the CLT, while retaining explicit, finite-N validation paths whenever exactness is needed.

We demonstrated how the exact/CLT toolkit informs sizing and inference in probabilistic data structures (Bloom merges; MinHash planning), incidence-graph coverage targets, cryptographic mechanisms and blockchain protocols that rely on random subset selection (e.g., threshold recovery via W_t, committee overlaps via Y, and redundancy profiles via Z_ℓ).

References

1. Barot, M., de la Peña, J.A.: Estimating the size of a union of random subsets of fixed cardinality. Elemente der Mathematik **56**(4), 163–169 (12 2001). https://doi.org/10.1007/PL00000552, https://ems.press/journals/em/articles/581
2. Beimel, A.: Secret-sharing schemes: A survey. In: Third international conference on Coding and cryptology. vol. 6639 LNCS, pp. 11–46 (2011). https://doi.org/10.1007/978-3-642-20901-7{_}2, https://www.researchgate.net/publication/220776045_Secret-Sharing_Schemes_A_Survey
3. Broder, A., Mitzenmacher, M.: Network applications of bloom filters: A survey. Internet Mathematics **1**(4), 485–509 (2004). https://doi.org/10.1080/15427951.2004.10129096;CTYPE:STRING:JOURNAL
4. Chatterjee, S., Meckes, E.: Multivariate normal approximation using exchangeable pairs. Latin American Journal of Probability and Mathematical Statistics **4**, 237–283 (1 2008), https://arxiv.org/pdf/math/0701464
5. Chung, N.C., Miasojedow, B.Z., Startek, M., Gambin, A.: Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data. BMC Bioinformatics **20**(15), 1–11 (12 2019). https://doi.org/10.1186/S12859-019-3118-5/FIGURES/6, https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3118-5
6. Cohen, E.: Min-Hash Sketches. Encyclopedia of Algorithms, Second Edition pp. 1282–1287 (1 2016). https://doi.org/10.1007/978-1-4939-2864-4{_}573, https://link.springer.com/rwe/10.1007/978-1-4939-2864-4_573
7. Kalinka, A.T.: The probability of drawing intersections: extending the hypergeometric distribution (5 2013), https://arxiv.org/pdf/1305.0717
8. McCormick, W.P., Lyons, N.I., Hutcheson, K.: Distributional properties of jaccard's index of similarity. Communications in Statistics - Theory and Methods **21**(1), 51–68 (1 1992). https://doi.org/10.1080/03610929208830764;PAGE:STRING:ARTICLE/CHAPTER, https://www.tandfonline.com/doi/abs/10.1080/03610929208830764

9. Real, R., Vargas, J.M.: The Probabilistic Basis of Jaccard's Index of Similarity. Systematic Biology **45**(3), 380–385 (9 1996). https://doi.org/10.1093/SYSBIO/45.3.380, https://dx.doi.org/10.1093/sysbio/45.3.380
10. Reinert, G., Röllin, A.: Multivariate normal approximation with Stein's method of exchangeable pairs under a general linearity condition. Annals of Probability **37**(6), 2150–2173 (12 2009). https://doi.org/10.1214/09-AOP467, http://arxiv.org/abs/0711.1082http://dx.doi.org/10.1214/09-AOP467
11. Shamir, A.: How to share a secret. Communications of the ACM **22**(11), 612–613 (11 1979). https://doi.org/10.1145/359168.359176, https://dl.acm.org/doi/10.1145/359168.359176

A Exact recursions for $U, I, (U, I)$

This section give a few exact recursion formula using result from section 4.

A.1 Univariate recursions for union (C_m)

Two parties. Let $m = 2$ and set $v_2 := n_1 + n_2 - u_2$. Then $|P_1 \cup P_2| = u_2$ iff $|P_1 \cap P_2| = v_2$. Conditioned on P_1,

$$\mathbb{P}(U = u_2) = \mathbb{P}(|P_1 \cap P_2| = v_2) = \frac{\binom{n_1}{v_2}\binom{N-n_1}{n_2-v_2}}{\binom{N}{n_2}}, \qquad v_2 \in \big\{\max(0, n_1+n_2-N), \ldots, \min(n_1, n_2)\big\}.$$

Equivalently, the global count of ordered pairs with union x is

$$C_2(u_2) = \binom{N}{n_1}\binom{n_1}{v_2}\binom{N-n_1}{n_2-v_2}, \quad \text{so} \quad \mathbb{P}(U = u_2) = \frac{C_2(u_2)}{\binom{N}{n_1}\binom{N}{n_2}}.$$

Three parties. Conditioned on the union of the first two parties, $u_2 := |P_1 \cup P_2|$. To reach final union u_3 after adding P_3: - pick exactly $t := n_3 + u_2 - u_3$ elements of P_3 from the previous union (so that P_3 contributes $u_3 - u_2$ new elements), and - pick $u_3 - u_2$ new elements from outside the previous union.

Given a specific previous union of size u_2, the number of ways to do this is $\binom{u_2}{t}\binom{N-u_2}{u_3-u_2}$. Summing over feasible u_2 and weighting by the count of (P_1, P_2) with union u_2,

$$C_3(u_3) = \sum_{u_2} \binom{u_2}{n_3 + u_2 - u_3}\binom{N-u_2}{u_3-u_2} C_2(u_2),$$

hence

$$\mathbb{P}(U = u_3) = \frac{C_3(u_3)}{\prod_{i=1}^{3}\binom{N}{n_i}}.$$

With the binomial-zero convention, the sum can be taken over all integers, but one convenient explicit range is

$$u_{2,\min} = \max\{n_1, n_2, u_3 - n_3,\, 0\}, \qquad u_{2,\max} = \min\{n_1 + n_2, N\}.$$

General m (global recursion). Define the global counts $C_m(u_m)$ of ordered m-tuples with union size u_1 by the base

$$C_1(u_1) = \begin{cases} \binom{N}{n_1}, & u_1 = n_1, \\ 0, & \text{otherwise}, \end{cases}$$

and the recursion, for $m \geq 2$,

$$C_m(u_m) = \sum_{u_{m-1}} \binom{u_{m-1}}{n_m + u_{m-1} - u_m}\binom{N-u_{m-1}}{u_m-u_{m-1}} C_{m-1}(u_{m-1}). \qquad (21)$$

A convenient feasible range is

$$u_{\min} = \max\{\, n_{m-1}^{\max},\, u_m - n_m,\, s_{m-1} - (m-2)N,\, 0 \,\}, \quad u_{\max} = \min\{\, u_m,\, N,\, s_{m-1} \,\},$$

where $s_{m-1} = \sum_{r=1}^{m-1} n_r$ and $n_{m-1}^{\max} = \max_{r \le m-1} n_r$. Finally,

$$\mathbb{P}(U = u_m) = \frac{C_m(u_m)}{\prod_{i=1}^{m} \binom{N}{n_i}}. \tag{22}$$

A.2 Univariate recursions for intersection (D_m)

Two parties. Conditioned on P_1, the intersection size is hypergeometric:

$$\mathbb{P}(I = v_2) = \frac{\binom{n_1}{v_2}\binom{N-n_1}{n_2-v_2}}{\binom{N}{n_2}}, \qquad v_2 \in \big\{\max(0, n_1 + n_2 - N), \ldots, \min(n_1, n_2)\big\}.$$

Equivalently, the global count is

$$D_2(v_2) = \binom{N}{n_1}\binom{n_1}{v_2}\binom{N-n_1}{n_2-v_2}, \quad \text{so} \quad \mathbb{P}(I = v_2) = \frac{D_2(v_2)}{\binom{N}{n_1}\binom{N}{n_2}}.$$

Three parties. Conditioned on the two-way intersection $v_2 := |P_1 \cap P_2|$. To have three-way intersection v_3, pick exactly v_3 of those v_2 common elements for P_3, and choose the remaining $n_3 - v_3$ elements of P_3 outside $P_1 \cap P_2$. This gives the factor $\binom{v_2}{v_3}\binom{N-v_2}{n_3-v_3}$. Summing over feasible v_2,

$$D_3(v_3) = \sum_{v_2} \binom{v_2}{v_3}\binom{N-v_2}{n_3-v_3} D_2(v_2), \qquad \mathbb{P}(I = v_3) = \frac{D_3(v_3)}{\prod_{i=1}^{3} \binom{N}{n_i}}.$$

A convenient explicit range is

$$v_{2,\min} = \max\{\, v_3,\, n_1 + n_2 - N,\, 0 \,\}, \qquad v_{2,\max} = \min\{\, n_1,\, n_2 \,\}.$$

(Outside this range the binomials vanish.)

General m (global recursion). Define global counts $D_m(v_m)$ of ordered m-tuples with intersection size v_m by the base

$$D_1(v_1) = \begin{cases} \binom{N}{n_1}, & v_1 = n_1, \\ 0, & \text{otherwise}, \end{cases}$$

and the recursion, for $m \ge 2$,

$$D_m(v_m) = \sum_{v_{m-1}} \binom{v_{m-1}}{v_m}\binom{N-v_{m-1}}{n_m-v_m} D_{m-1}(v_{m-1}). \tag{23}$$

A convenient feasible range is

$$v_{m-1,\min} = \max\{\, v_m,\, \textstyle\sum_{r=1}^{m-1} n_r - (m-2)N,\, 0 \,\}, \quad v_{m-1,\max} = \min\{\, n_{m-1}^{\min} \,\},$$

where $n_{m-1}^{\min} = \min_{r \le m-1} n_r$. Finally,

$$\mathbb{P}(I = v_m) = \frac{D_m(v_m)}{\prod_{i=1}^{m} \binom{N}{n_i}}. \tag{24}$$

From standard feasibility constraints,

$$\max\Big\{0,\ \sum_{r=1}^{m} n_r - (m-1)N\Big\} \ \le\ v_m \ \le\ \min_{r \le m} n_r, \quad \max_{r \le m} n_r \ \le\ u_m \ \le\ N.$$

A.3 Bivariate recursion (F_m)

For $k \ge 1$, write $S_k = (n_1, \ldots, n_k)$ and define the count

$$F_k(u_k, v_k) := \Big|\Big\{(P_1, \ldots, P_k) :\ \big|\cup_{r \le k} P_r\big| = u_k,\ \big|\cap_{r \le k} P_r\big| = v_k\Big\}\Big|.$$

Then the joint PMF of (U, I) is

$$\mathbb{P}(U = u_m,\ I = v_m) = p_{U,I}(u_m, v_m) = \frac{F_m(u_m, v_m)}{\prod_{r=1}^{m} \binom{N}{n_r}}, \tag{25}$$

for (u_m, v_m) in the feasible region $\mathcal{R}$ (below), where F_m is defined recursively as follows.

Base case ($m = 1$).

$$F_1(u_1, v_1) = \mathbf{1}\{u_1 = v_1 = n_1\} \binom{N}{n_1}$$

Two parties ($m = 2$).

$$F_2(u_2, v_2) = \mathbf{1}\{u_2 = n_1 + n_2 - v_2\} \binom{N}{n_1} \binom{n_1}{v_2} \binom{N - n_1}{n_2 - v_2}. \tag{26}$$

General recursion ($m \ge 3$). Let

$$s_{m-1} := \sum_{r=1}^{m-1} n_r, \qquad n_{m-1}^{\min} := \min_{r \le m-1} n_r, \qquad n_{m-1}^{\max} := \max_{r \le m-1} n_r.$$

Define bounds

$$\begin{aligned} a_{m-1} &:= \max\{\, v_m, s_{m-1} - (m-2)N \,\}, \\ b_{m-1} &:= n_{m-1}^{\min}, \\ c_{m-1}(v_{m-1}) &:= \max\Big\{\, n_{m-1}^{\max}, \Big\lceil \frac{s_{m-1} - v_{m-1}}{m-2} \Big\rceil \Big\}, \\ d_{m-1}(v_{m-1}) &:= s_{m-1} - (m-2)\, v_{m-1}. \end{aligned}$$

Then for all $m \geq 3$,

$$F_m(u_m, v_m) = \sum_{v_{m-1}=a_{m-1}}^{b_{m-1}} \sum_{u_{m-1}=c_{m-1}(v_{m-1})}^{d_{m-1}(v_{m-1})} \binom{v_{m-1}}{v_m} \binom{u_{m-1} - v_{m-1}}{n_m + u_{m-1} - u_m - v_m} \binom{N - u_{m-1}}{u_m - u_{m-1}} \cdot F_{m-1}(u_{m-1}, v_{m-1}). \tag{27}$$

Feasible region.

$$\mathcal{R} = \Big\{ (u_m, v_m) \in \mathbb{Z}^2 : s_m - (m-2)N \leq v_m \leq n_m^{\min}, \quad \max\{n_m^{\max}, \lceil \frac{s_m - v_m}{m-2} \rceil\} \leq u_m \leq s_m - (m-2)v_m \Big\} \tag{28}$$

Proof. Condition on (u_{m-1}, v_{m-1}) from the first $m-1$ parties. To add party m, (i) keep v_m elements in the intersection: $\binom{v_{m-1}}{v_m}$; (ii) choose $s := n_m + u_{m-1} - u_m - v_m$ elements from the "union-only" band $u_{m-1} - v_{m-1}$: $\binom{u_{m-1}-v_{m-1}}{s}$; and (iii) add $t := u_m - u_{m-1}$ new elements outside the previous union: $\binom{N-u_{m-1}}{t}$. Note $n_m = v_m + s + t$. The bounds $a_{m-1}, b_{m-1}, c_{m-1}, d_{m-1}$ are exactly those ensuring $0 \leq v_m \leq v_{m-1}$, $0 \leq s \leq u_{m-1} - v_{m-1}$, $0 \leq t \leq N - u_{m-1}$, and feasibility for the first $m-1$ parties. Summing over feasible (u_{m-1}, v_{m-1}) yields (27); normalising by $\prod_r \binom{N}{n_r}$ gives (25).

B Moments for the level counts Z_k

We derive $\mathbb{E}(Z_k)$, $\mathrm{Var}(Z_a)$, and $\mathrm{Cov}(Z_a, Z_b)$ starting from single- and two-item probabilities.

One-item probabilities and means. For a single item i and $\ell \in \{0, \ldots, m\}$,

$$p_\ell(\alpha) = \mathbb{P}(R(i) = \ell) = [t^\ell] \prod_{r=1}^{m} \big((1 - \alpha_r) + \alpha_r t\big).$$

Therefore $\mathbb{E}(Z_\ell) = N p_k(\alpha)$.

Two-item probabilities. Fix distinct items $u \neq v$. For party r the pair $(\mathbf{1}\{u \in P_r\}, \mathbf{1}\{v \in P_r\})$ has probabilities

$$\pi_{00}^{(r)} = \frac{(N - n_r)(N - n_r - 1)}{N(N-1)}, \quad \pi_{10}^{(r)} = \pi_{01}^{(r)} = \frac{n_r(N - n_r)}{N(N-1)}, \quad \pi_{11}^{(r)} = \frac{n_r(n_r - 1)}{N(N-1)}.$$

Define $\phi_r(z, w) := \pi_{00}^{(r)} + \pi_{10}^{(r)} z + \pi_{01}^{(r)} w + \pi_{11}^{(r)} zw$. Across parties,

$$q_{a,b} \;:=\; \Pr\big(R(u) = a,\ R(v) = b\big) \;=\; [z^a w^b] \prod_{r=1}^{m} \phi_r(z, w).$$

Finite-N variance–covariance identities. Recall $Z_\ell = \sum_{i=1}^{N} I_i^{(\ell)}$. Then

$$\mathrm{Var}(Z_a) = \sum_{i=1}^{N} \mathrm{Var}(I_i^{(a)}) + \sum_{u \neq v} \mathrm{Cov}(I_u^{(a)}, I_v^{(a)}) = N\,p_a(1-p_a) + N(N-1)\left(q_{a,a} - p_a^2\right),$$

$$\mathrm{Cov}(Z_a, Z_b) = \sum_{u \neq v} \mathrm{Cov}(I_u^{(a)}, I_v^{(b)}) \;-\; N\,p_a p_b = N(N-1)\,q_{a,b} - N^2 p_a p_b, \qquad a \neq b.$$

These are exact for every N.

Asymptotic expansion of $q_{a,b}$. Write $\alpha_r = n_r/N$. Expand each ϕ_r to first order in $(N-1)^{-1}$:

$$\phi_r(z,w) = \big((1-\alpha_r) + \alpha_r z\big)\big((1-\alpha_r) + \alpha_r w\big) + \frac{\alpha_r(1-\alpha_r)}{N-1}\,(-1 + z + w - zw).$$

Multiplying over r and extracting coefficients,

$$q_{a,b} \;=\; p_a p_b \;+\; \frac{1}{N-1}\,S_{a,b}(\alpha) \;+\; O\big((N-1)^{-2}\big),$$

where

$$S_{a,b}(\alpha) := \sum_{r=1}^{m} \alpha_r(1-\alpha_r)\big\{p_{a-1}^{(-r)} p_b^{(-r)} + p_a^{(-r)} p_{b-1}^{(-r)} - p_a^{(-r)} p_b^{(-r)} - p_{a-1}^{(-r)} p_{b-1}^{(-r)}\big\}. \quad (29)$$

Here $p_\ell^{(-r)} := [t^\ell] \prod_{s \neq r}((1-\alpha_s) + \alpha_s t)$ and $p_{-1}^{(-r)} := 0$. Equivalently, in forward-difference form,

$$S_{a,b}(\alpha) \;=\; -\sum_{r=1}^{m} \alpha_r(1-\alpha_r)\,\Delta_a^{(-r)}\,\Delta_b^{(-r)}, \qquad \Delta_\ell^{(-r)} := p_\ell^{(-r)} - p_{\ell-1}^{(-r)}.$$

Leading-order covariance. Substitute $q_{a,b}$ into the finite-N identities to obtain

$$\Sigma_N \;=\; N\,\Sigma(\alpha) \;+\; O(1),$$

with entries

$$\Sigma_{aa}(\alpha) = p_a(1-p_a) - \sum_{r=1}^{m} \alpha_r(1-\alpha_r)\big(\Delta_a^{(-r)}\big)^2, \qquad \Sigma_{ab}(\alpha) = -p_a p_b - \sum_{r=1}^{m} \alpha_r(1-\alpha_r)\Delta_a^{(-r)}\Delta_b^{(-r)}\ (a \neq b).$$

The correction term is negative semidefinite (a sum of outer products with a minus sign); this reflects the weak negative dependence induced by sampling without replacement.

Sanity checks. (i) If parties included each item independently with probabilities α_r (*with* replacement), then $\phi_r(z,w) = ((1-\alpha_r)+\alpha_r z)((1-\alpha_r)+\alpha_r w)$, $S_{a,b} \equiv 0$, and $\Sigma(\alpha) = \mathrm{diag}(p) - pp^\top$ (multinomial covariance).
(ii) In the ultra-sparse regime where $p_\ell \to 0$ and $Np_\ell \to \lambda \in (0,\infty)$, the count Z_ℓ converges in law to Poisson(λ).

C Stein proof of the level-count CLT

This appendix proves Theorem 3 in full detail. We first explain the swap coupling; then we verify the three standard Stein ingredients: *regression*, *bounded increments*, and *matching second moments*. Finally, we invoke the multivariate Stein bound.

The swap coupling (preserves $|P_r| = n_r$) Given a configuration $(P_1, \dots, P_m)$, construct $(P'_1, \dots, P'_m)$ as follows.

1. Pick an index $K \sim \mathrm{Unif}([N])$.
2. For each party $r = 1, \dots, m$ *independently of other parties*:
   - If $K \in P_r$, choose $U_r \in [N] \setminus P_r$ uniformly and swap $K \leftrightarrow U_r$ (remove K from P_r, add U_r).
   - If $K \notin P_r$, choose $V_r \in P_r$ uniformly and swap $K \leftrightarrow V_r$ (add K to P_r, remove V_r).

Each swap preserves $|P_r| = n_r$, so the joint law of $(P'_1, \dots, P'_m)$ is again uniform over the same state space. The mapping is an involution (doing it twice returns to the starting configuration), so the pair

$$\mathbf{Z} := \big(Z_0, \dots, Z_m\big), \qquad \mathbf{Z}' := \big(Z'_0, \dots, Z'_m\big)$$

is *exchangeable*. Define the increment $\Delta := \mathbf{Z}' - \mathbf{Z}$.

How many coordinates can change? Only the tagged item K and, for each party, at most one partner (U_r or V_r) can change their coverage $R(\cdot)$ by ± 1. Hence at most $2m$ items change level. When a single item changes its coverage by ± 1, the level counts move one unit between two *adjacent* bins. Therefore, for each k, $\Delta_k \in \{-2m, \dots, 2m\}$ and, more sharply,

$$\|\Delta\|_1 \le 2m, \qquad \|\Delta\|_2 \le 2\sqrt{m}.$$

These deterministic bounds yield moment bounds after scaling by $N^{-1/2}$.

The regression property Write $\bar{\mathbf{Z}} := \mathbf{Z} - \mathbb{E}(\mathbf{Z})$ and let $\mathsf{T} = \{x : \mathbf{1}^\top x = 0\}$ be the tangent subspace, with projector Π onto T. We prove

$$\mathbb{E}\big[\Delta \mid \mathbf{Z}\big] = -\frac{1}{N} \Pi\, \bar{\mathbf{Z}}. \tag{30}$$

Step 1: contribution from the tagged item K. Condition on the entire configuration. The level of K *before* the swap, say $R(K) = a$, contributes $-e_a$ to Δ if K leaves level a, where e_a is the a-th standard basis vector in $\mathbb{R}^{m+1}$. *After* the full set of party swaps, the coverage of K is re-drawn from its *stationary one-item law* given the rest, whose mean indicator vector is the population average $\mathbb{E}(\mathbf{Z})/N$. Therefore, the conditional expectation of the *new* level vector for K equals $(\mathbb{E}(\mathbf{Z}))/N$. Averaging over $K \sim \mathrm{Unif}([N])$ and using $\mathbb{P}(R(K) = a \mid Z) = Z_a/N$ (exchangeability of items), we get

$$\mathbb{E}\big[(\text{change caused by } K) \;\big|\; \mathbf{Z}\big] = \frac{\mathbb{E}(\mathbf{Z})}{N} - \frac{\mathbf{Z}}{N} = -\frac{1}{N}\,\bar{\mathbf{Z}}.$$

Step 2: contributions from partner items. For a fixed party r, the partner (U_r or V_r) is chosen uniformly from the complement or from P_r respectively. Conditioned on K and the sets P_r, the two cases are symmetric and have opposite signs in expectation. Therefore, their *net* conditional drift is 0. Summing over parties yields zero expected contribution from all partners. Combining with Step 1 gives (30). Since $\mathbf{1}^\top \Delta = 0$ deterministically, we may insert Π without changing the equality.

Bounded increments and third moments Define the scaled, projected statistic

$$W \;:=\; \frac{1}{\sqrt{N}}\,(\Pi\Sigma(\alpha)\Pi^\top)^{-1/2}\,\Pi\,\bar{\mathbf{Z}}, \qquad W' \;:=\; W + \frac{1}{\sqrt{N}}\,(\Pi\Sigma(\alpha)\Pi^\top)^{-1/2}\,\Pi\,\Delta.$$

Using $\|\Pi\| \le 1$, $\|(\Pi\Sigma(\alpha)\Pi^\top)^{-1/2}\| = O(1)$ (fixed m, fixed α), and $\|\Delta\|_2 \le 2\sqrt{m}$,

$$\|W' - W\| \;\le\; C\,\frac{\|\Delta\|_2}{\sqrt{N}} \;\le\; \frac{C'}{\sqrt{N}}, \qquad \mathbb{E}\|W' - W\|^3 \;=\; O\big(N^{-3/2}\big).$$

Conditional second moments match $\Sigma(\alpha)$ Let $V := \mathbb{E}[\Delta\Delta^\top \mid Z]$ and $V_0 := \mathbb{E}V$. We claim

$$\frac{1}{N}\,V_0 \;=\; \Pi\,\Sigma(\alpha)\,\Pi^\top, \qquad \mathbb{E}\big\|V - V_0\big\|_{\mathrm{HS}} \;=\; O\big(N^{-1}\big). \tag{31}$$

Intuition: each swap touches only $O(1)$ items, so the change in quadratic forms depends only on $O(1)$ coordinates of $\mathbf{Z}$, leading to $O(1/N)$ fluctuations.

Derivation of V_0. Pick K uniformly. The increment Δ is a sum of at most $2m$ elementary moves that transfer one unit between adjacent levels. Averaging over the random positions of these moves and over the random K, the expected outer product $\mathbb{E}[\Delta\Delta^\top]$ reproduces the *per-item* contributions to the covariance of Z described in §5.3, scaled by $1/N$. Carrying out the coefficient extraction (Appendix B) gives the first relation in (31). The second relation follows because Δ depends on Z only through the levels of at most $2m$ items and the choice of K, giving $O(1/N)$ variability in conditional second moments.

Conclusion via multivariate Stein We now apply the exchangeable-pairs version of multivariate Stein's method: if (W, W') is exchangeable, $\mathbb{E}[W' - W \mid W] = -\lambda W + R$ with $\lambda = 1/N$ and $\|R\|$ negligible, and if the conditional second moments match the target covariance up to $O(1/N)$ while $\mathbb{E}\|W' - W\|^3 = O(N^{-3/2})$, then for smooth test functions h,

$$\big|\mathbb{E}(h(W)) - \mathbb{E}(h(Z))\big| \;\le\; \frac{C}{\sqrt{N}},$$

where $Z \sim \mathcal{N}(0, I_m)$ and C depends only on m and α. A standard smoothing step transfers this bound to the distance induced by indicators of convex sets. This yields Theorem 3. □

D Numerical validation

All figures in this section use the *exact* finite–N counts from the paper: the bivariate recursion F_m in A.3,Equation 27 for (U, I) and its univariate specialisations C_m (unions) and D_m (intersections) from Appendix A.1,A.2. Gaussian overlays are parameterised by the closed-form moments from §6.2, and for Jaccard by the delta–method (μ_J, σ_J^2) in §6.3. Where a continuous normal density is drawn over a discrete PMF, we apply a half-cell continuity correction.

D.1 Univariate distribution

The univariate plots use the global recursions C_m (union) and D_m (intersection) from Appendix A, which are the marginals of F_m (25). We show (i) exact PMFs, (ii) tail probabilities relevant for design, and (iii) large–N Gaussian overlays.

Union. Figure 1 shows the exact PMF $\mathbb{P}(U = u)$ for $N = 10$ and $S_3 = (3, 4, 5)$. Mass at $u = N$ corresponds to full coverage (cf. the recovery event in §7); the mode lies close to μ_U from §6.2.

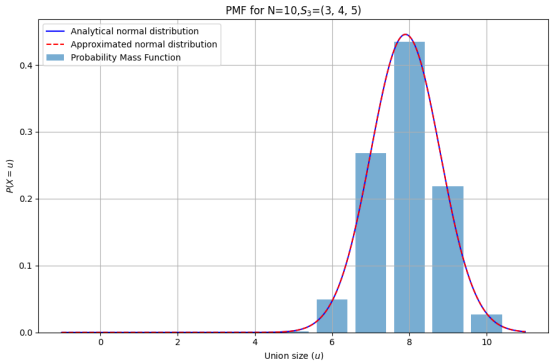


Fig. 1. Exact PMF of U for $N = 10$ and $S_3 = (3, 4, 5)$ (via recursion C_3).

For tail requirements in §7, we evaluate

$$\mathbb{P}(U \geq k) = \sum_{u=k}^{N} \frac{C_m(u)}{\prod_{r=1}^{m} \binom{N}{n_r}}, \tag{32}$$

displayed in Figure 2 for the same parameters.

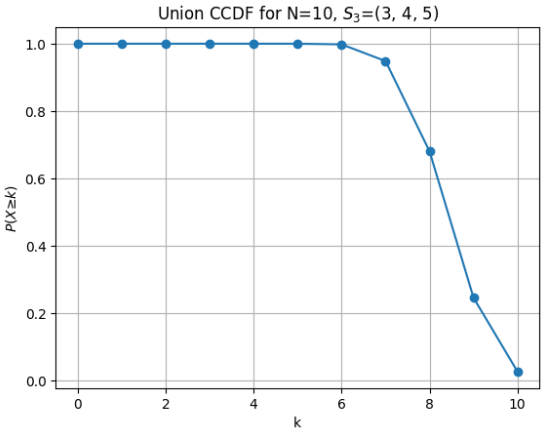
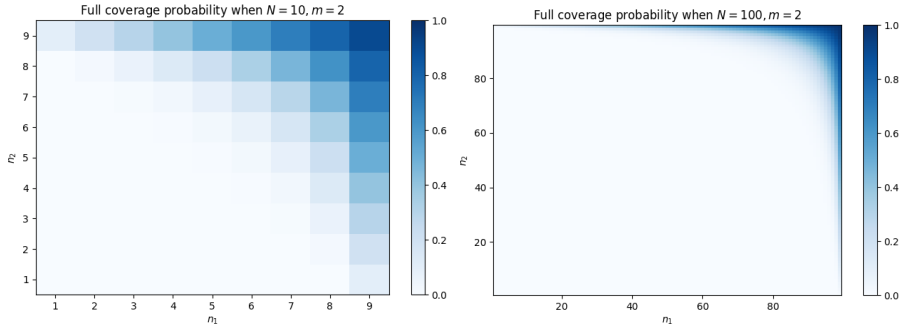


Fig. 2. Tail probability $\mathbb{P}(U \geq k)$ for $N = 10$ and $S_3 = (3, 4, 5)$ (via C_3).

The full-coverage probability is

$$\mathbb{P}(U = N) = \frac{C_m(N)}{\prod_{r=1}^{m} \binom{N}{n_r}},$$

and its dependence on (n_1, n_2) for $m = 2$ is summarised in Figure 3 (useful for merged-filter sizing in §7).



(a) $N = 10$ (b) $N = 100$

Fig. 3. Full-coverage probability $\mathbb{P}(U = N)$ for $m = 2$ across (n_1, n_2) (via C_2).

Intersection. Similarly, Figure 4 shows the exact PMF $\mathbb{P}(I = v)$ for $N = 10$ and $S_3 = (3, 4, 5)$.

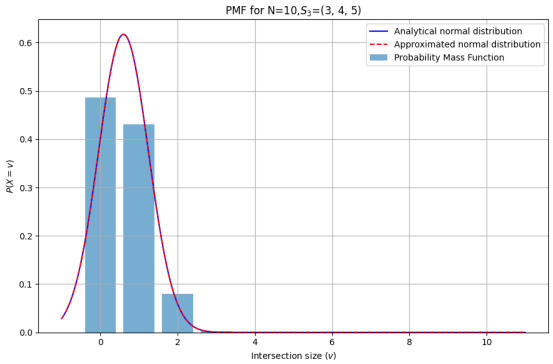


Fig. 4. Exact PMF of I for $N = 10$ and $S_3 = (3, 4, 5)$ (via recursion D_3).

Large–N behaviour. Theorem 3 gives an $O(N^{-1/2})$ quantitative CLT for (U, I) after standardisation, univariate normal approximations follow. Figures 5 and 6 overlay $\mathcal{N}(\mu_U, \sigma_U^2)$ and $\mathcal{N}(\mu_I, \sigma_I^2)$ (moments from §6.2) on the exact PMFs. Agreement is excellent away from the boundaries $u \approx N$ and $v \approx 0$ or $\min_r n_r$; in extremely sparse-overlap regimes (very small p_I) a Poisson or compound-Poisson surrogate for I can be sharper (§7).

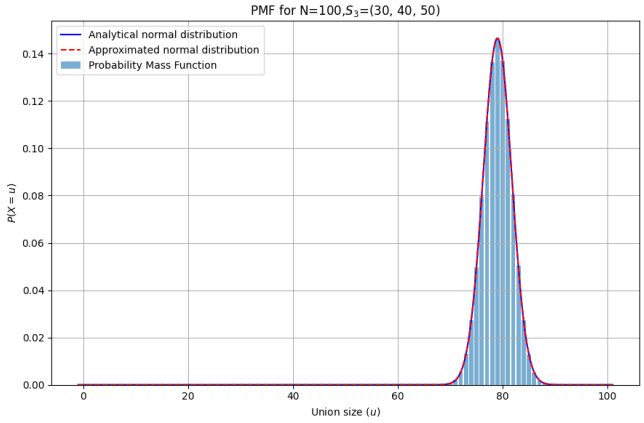


Fig. 5. Exact PMF of U for $N = 100$ and $S_3 = (30, 40, 50)$ with normal overlay $\mathcal{N}(\mu_U, \sigma_U^2)$; continuity correction applied.

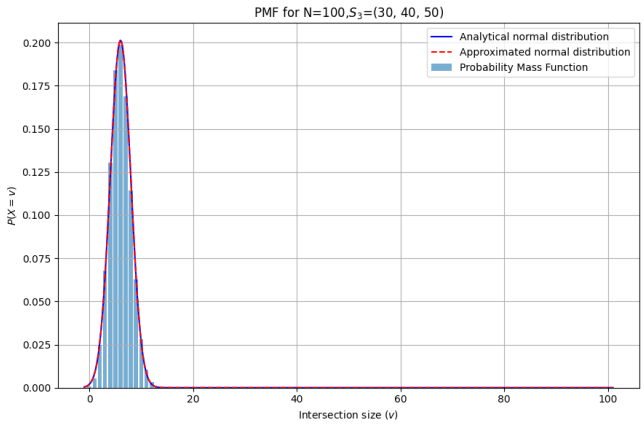


Fig. 6. Exact PMF of I for $N = 100$ and $S_3 = (30, 40, 50)$ with normal overlay $\mathcal{N}(\mu_I, \sigma_I^2)$; continuity correction applied.

D.2 Bivariate distribution

Figure 7 shows the *exact* joint PMF $p_{U,I}$ from (25) (via F_m) as a heat map, with isocontours of the Gaussian surrogate $\mathcal{N}_2(\mu, \Sigma)$ overlaid. The orientation and eccentricity of the ellipses are determined by the covariance $\mathrm{Cov}(U, I)$ in §6.2; the close alignment of isocontours with PMF level sets empirically supports the approximation guaranteed by the bivariate CLT, which provides an $O(N^{-1/2})$ finite–N error for the standardised vector. As expected, departures are most visible (not shown) when (u, v) sits near the boundary of the feasible region $\mathcal{R}$ (defined in Appendix A.3), where lattice effects and truncation become non-negligible.

D.3 Jaccard distribution

We validate both the *exact* Jaccard constructions from §6.3 and the delta–method surrogate.

Expected value over (n_1, n_2). Panel (a) of Figure 8 plots the exact $\mathbb{E}[J_N]$ for $m = 2$, obtained by summing the hypergeometric law of $I = |P_1 \cap P_2|$ (Proposition 1). Panel (b) shows the delta–method mean μ_J derived from (μ_N, Σ_N); panel (c) reports the relative error.

Empirical histogram vs two normal fits. Figure 9 compares the *empirical* distribution of J_N (histogram from Monte Carlo draws of $(P_1, \ldots, P_m)$ under fixed–cardinality sampling) with two normal curves: (i) a moment–fit normal using the empirical mean/SD of J_N; and (ii) the *delta–method* normal $\mathcal{N}(\mu_J, \sigma_J^2)$ computed from (μ_N, Σ_N) in §6.2. As N increases, the normal curves track the histogram increasingly closely and the distribution narrows, consistent with the $N^{-1/2}$ scaling predicted by the CLT and delta method.

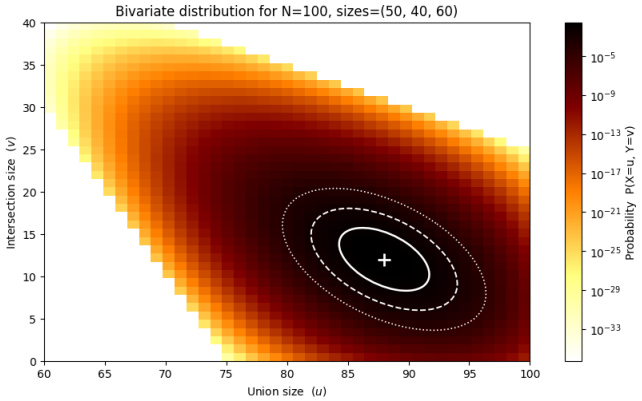
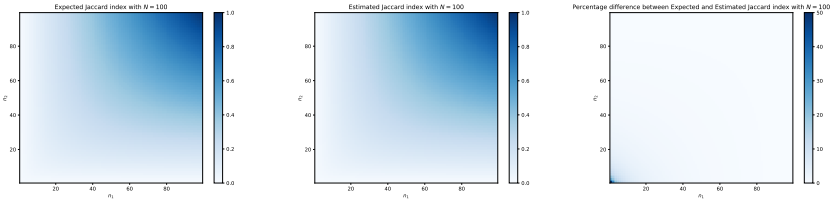
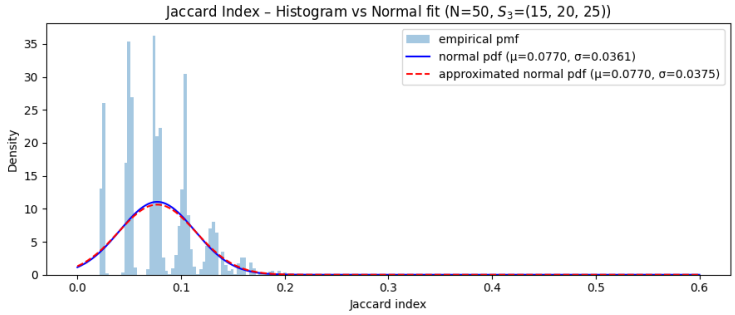


Fig. 7. Exact joint PMF of (U, I) (heat map via F_m) with $\mathcal{N}_2(\mu_N, \Sigma_N)$ isocontours (moments from §6.2) overlaid; continuity correction applied.

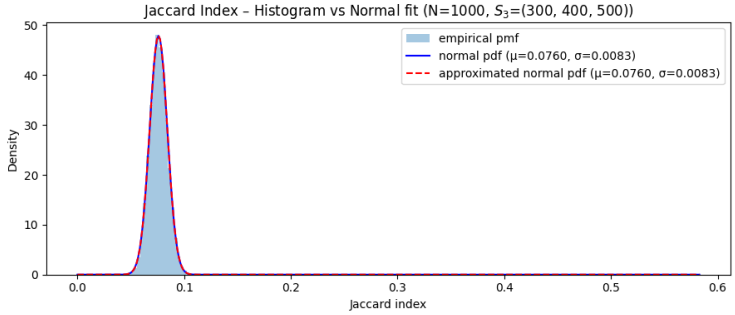


(a) Exact $\mathbb{E}[J_N]$ (Prop. 1) (b) Delta–method μ_J (§6.3)(c) $100\times|\mathbb{E}[J_N]-\mu_J|/\mathbb{E}[J_N]$

Fig. 8. Expected Jaccard index across (n_1, n_2) with $N = 100$. Exact (a) vs delta–method (b); relative error (c).



(a) $N = 50$.



(b) $N = 1000$.

Fig. 9. Exact PMF of J_N (via Prop. 1) for a representative parameter choice versus the delta–method $\mathcal{N}(\mu_J, \sigma_J^2)$. Continuity correction and truncation to $[0, 1]$ applied to the overlay.