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

Alice K. Ng, Jiahua Xu, Paolo Tasca

August 19, 2025

Abstract

We develop a distributional theory for coverage and overlap in random set systems under fixed–cardinality, without–replacement sampling. First, we give an exact, efficiently memoised recursion for the joint distribution of the union and intersection sizes of m subsets; the univariate laws follow by marginalisation. Second, using Stein's method with a swap–based exchangeable–pairs coupling, we prove a two–dimensional central limit theorem with an $O(N^{-1/2})$ error bound, from which the usual univariate normal approximations follow. Third, for the Jaccard index we obtain a general construction for arbitrary m induced by the joint distribution, and a Gaussian approximation via the delta method with closed–form mean and variance.

The exact recursion enables computation on moderate instances, while the Gaussian surrogates are accurate across broader regimes. We illustrate the utility of these results for the design and analysis of probabilistic data structures, incidence–graph models, and secret–sharing schemes.

Contents

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 i (fixed–cardinality, without–replacement sampling). In many applications one needs not only expectations but the full distributions and joint behaviour of

$$X_N = \Big|\bigcup_{r=1}^{m} P_r\Big| \quad \text{(total coverage)}, \qquad Y_N = \Big|\bigcap_{r=1}^{m} P_r\Big| \quad \text{(full overlap)},$$

as well as functionals such as the Jaccard similarity $J_N = Y_N/X_N$. While the *univariate* laws for union or intersection sizes are well studied, general bivariate laws and scalable approximations tailored to the without–replacement model (with arbitrary $n_1, \ldots, n_m$) are less developed.

We study this fixed–size model with arbitrary $\{n_r\}$. Our aims are twofold: (i) to give exact, computable finite–N laws when feasible; and (ii) to provide accurate normal approximations for the bivariate (and hence univariate) distributions with explicit error control for finite and large N. The results unify the analysis of X_N, Y_N, (X_N, Y_N), and J_N, and can be used directly for design and inference in systems that aggregate or compare sets.

Contributions.

- **Exact bivariate law (general m, arbitrary $\{n_r\}$).** We derive a recursion $F_m(u, v; S_m)$, with $S_m = (n_1, \ldots, n_m)$, that counts ordered m-tuples $(P_1, \ldots, P_m)$ with $|\cup_r P_r| = u$ and $|\cap_r P_r| = v$, together with sharp feasibility bounds for (u, v). This yields the exact joint pmf $p_{X_N,Y_N}(u, v)$, and the marginal laws follow by summation:

$$\mathbb{P}(X_N = u) = \sum_v p_{X_N,Y_N}(u, v), \qquad \mathbb{P}(Y_N = v) = \sum_u p_{X_N,Y_N}(u, v).$$

- **Moments and a bivariate CLT.** Using an indicator-variable calculus we obtain closed forms for $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$ and $\mathrm{Cov}(X_N, Y_N)$. Via an exchangeable-pairs Stein coupling, we prove a bivariate normal approximation for (X_N, Y_N) with an $O(N^{-1/2})$ error rate in Wasserstein distance. The univariate normal approximations for X_N and Y_N follow immediately by marginalisation (with the same finite–N rate).
- **Jaccard index.** For $m \geq 2$, we obtain the exact finite–N distribution of $J_N = Y_N/X_N$ from the bivariate counts F_m. We also give a delta–method Gaussian approximation using the derived μ_X, μ_Y and Σ_N, yielding closed-form mean and variance for J_N.

- **Algorithms, validations, and use cases.** We detail a practical memoised evaluator for F_m, and we validate the Gaussian and delta–method approximations against exact pmfs across regimes. We then show how these tools inform probabilistic sizing for merged Bloom filters, Jaccard inference for MinHash, threshold recovery/leakage analyses in secret sharing, and coverage targets in incidence-graph models.

2 Related work

Random set systems and union/intersection statistics. Classical results describe the *univariate* distribution and moments of the size of a union when s subsets of equal size k are drawn uniformly without replacement from a ground set of size N; see, e.g., Barot and de la Peña [1], who give a closed form for $\mathbb{P}\big(\big|\bigcup_{r=1}^{s} P_r\big| = u\big)$ together with exact expressions for the mean and variance; Kalinka [6], who give a closed form for $\mathbb{P}\big(\big|\bigcap_{r=1}^{s} P_r\big| = u\big)$. Much of this literature treats unions alone and assumes equal subset sizes; explicit results for intersections typically appear only as marginals or via inclusion–exclusion, and *joint* laws for $\big(\big|\bigcup_r P_r\big|, \big|\bigcap_r P_r\big|\big)$ are not standard. Our work gives a finite–N recursion for the joint distribution that accommodates arbitrary cardinalities $\{n_r\}$ and provides sharp feasibility bounds, from which the univariate laws follow by marginalisation.

Bernoulli presence models and the Jaccard index. A distinct line of work analyses the Jaccard index under independent Bernoulli presence/absence models (each item present in each set with some probability), yielding exact null distributions and efficient approximations for hypothesis testing. These results are not directly applicable to our fixed–cardinality, without–replacement model, where the relevant combinatorics are hypergeometric rather than binomial. We bridge this gap by providing (i) an exact finite–N distribution for $m = 2$, (ii) a construction for general m via the joint law of union and intersection, and (iii) a principled Gaussian surrogate for J_N via the delta method with closed–form moments.

Stein's method and multivariate normal approximation. Stein's method offers quantitative normal approximation under dependence. Multivariate versions via exchangeable pairs were developed by Chatterjee–Meckes [4] and by Reinert–Röllin [7] under a general linearity condition. We instantiate these techniques in the fixed–cardinality random set model using a swap–based exchangeable–pairs coupling, yielding a bivariate CLT for (X_N, Y_N) with an explicit $O(N^{-1/2})$ error bound and, as corollaries, the usual univariate normal approximations.

Probabilistic data structures and sketch design. Bloom's filter and the MinHash framework motivate distributional questions about unions, intersections, and ratios such as Jaccard. Prior analyses often rely on independence or Poissonisation heuristics; our exact recursion and Gaussian limits provide finite–N design rules for merged filters and similarity sketches under without–replacement sampling. Cardinality sketches such as HyperLogLog address distinct problems (estimating $\left|\bigcup_r P_r\right|$ from hashed streams) but likewise illustrate the value of explicit distributional control.

Intersection/coverage viewpoints in random graph models. Random intersection graphs build edges from overlaps of random attribute sets assigned to vertices. That literature characterises degrees, clustering, and phase transitions under various sampling schemes; however, exact finite–N joint laws for coverage and overlap of *fixed* families of sets are typically not the focus. Our counts and CLT supply inputs that can be repurposed for incidence–graph design questions where set sizes are prescribed.

Summary of differences. In contrast to (i) univariate union results with equal subset sizes, (ii) Jaccard analyses under Bernoulli presence, and (iii) asymptotic graph–level properties, we provide (a) a finite–N, computable recursion for the *joint* $\left(\left|\bigcup_r P_r\right|, \left|\bigcap_r P_r\right|\right)$ law for arbitrary $\{n_r\}$, (b) a multivariate Stein CLT with a quantified $O(N^{-1/2})$ rate for the fixed–cardinality model, and (c) exact and approximate laws for J_N induced by (a)–(b).

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).

Our primary statistics are

$$X_N = \left|\bigcup_{r=1}^{m} P_r\right| \quad \text{(coverage)}, \qquad Y_N = \left|\bigcap_{r=1}^{m} P_r\right| \quad \text{(full overlap)}.$$

We denote the joint law of (X_N, Y_N) by

$$p_{X,Y}(u,v) := \mathbb{P}(X_N = u,\, Y_N = v),$$

with support restricted to the feasible region for (u,v) determined by $(N; n_1, \ldots, n_m)$ (made explicit in §3.1). For brevity, write $S_m = (n_1, \ldots, n_m)$ and $\alpha = (\alpha_1, \ldots, \alpha_m)$.

Remarks. *Asymptotics and rounding.* In statements that take $N \to \infty$ with $\alpha_r \in (0,1)$ fixed, any choice of integer rounding for $n_r(N)$ with $n_r/N \to \alpha_r$ (e.g., $n_r = \lfloor \alpha_r N \rfloor$) yields identical limits; our finite–N formulas use the exact integers n_r.

4 Bivariate distribution

4.1 Exact bivariate pmf: a combinatorial recursion

Theorem 4.1. *For $k \geq 1$, write $S_k = (n_1, \ldots, n_k)$ and define the count*

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

Then the joint pmf of (X_N, Y_N) is

$$\mathbb{P}(X_N = u_m, \ Y_N = v_m) = p_{X,Y}(u_m, v_m) = \frac{F_m(u_m, v_m; S_m)}{\prod_{r=1}^{m} \binom{N}{n_r}}, \tag{4.1}$$

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

Base case ($m = 1$).

$$F_1(u_1, v_1; S_1) = \mathbf{1}\{u_1 = v_1 = n_1\} \binom{N}{n_1}$$

Two parties ($m = 2$).

$$F_2(u_2, v_2; S_2) = \mathbf{1}\{u_2 = n_1 + n_2 - v_2\} \binom{N}{n_1} \binom{n_1}{v_2} \binom{N - n_1}{n_2 - v_2}. \tag{4.2}$$

General recursion ($m \geq 3$). *Let*

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

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 $m \geq 3$,

$$\begin{aligned} F_m(u_m, v_m; S_m) = &\sum_{v_{m-1}=a_{m-1}}^{b_{m-1}} \ \sum_{u_{m-1}=c_{m-1}(v_{m-1})}^{d_{m-1}(v_{m-1})} \binom{v_{m-1}}{v_m} \binom{u_{m-1} - v_{m-1}}{n_m + u_{m-1} - u_m - v_m} \binom{N - u_{m-1}}{u_m - u_{m-1}} \\ &\cdot F_{m-1}(u_{m-1}, v_{m-1}; S_{m-1}). \end{aligned} \tag{4.3}$$

Feasible region.

$$\begin{aligned}\mathcal{R} = \Big\{(u_m, v_m) \in \mathbb{Z}^2 : s_m - (m-2)N \le v_m \le n_m^{\min},\\ \max\{n_m^{\max}, \lceil \frac{s_m - v_m}{m-2} \rceil\} \le u_m \le s_m - (m-2)v_m\Big\}\end{aligned} \tag{4.4}$$

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 $0 \le v_m \le v_{m-1}$, $0 \le s \le u_{m-1} - v_{m-1}$, $0 \le t \le N - u_{m-1}$, and feasibility for the first $m-1$ parties. Summing over feasible (u_{m-1}, v_{m-1}) yields (4.3); normalising by $\prod_r \binom{N}{n_r}$ gives (4.1). □

4.2 Moments via indicator variables

Let $\alpha_r := n_r/N$. For a *fixed* item, $\mathbb{P}(i \in P_r) = \alpha_r$ independently across r. Hence

$$\mathbb{E}(X_N) = \mu_X = N\Big(1 - \prod_{r=1}^{m}(1-\alpha_r)\Big) =: N\, p_X,$$

$$\mathbb{E}(Y_N) = \mu_Y = N\prod_{r=1}^{m}\alpha_r =: N\, p_Y.$$

Exact finite-N formulas (and their $N \to \infty$ expansions) for $\mathrm{Var}(X_N)$, $\mathrm{Var}(Y_N)$ and $\mathrm{Cov}(X_N, Y_N)$ follow from a standard indicator calculus with without-replacement corrections; we record the closed forms here and defer algebra to Appendix A:

$$\mathrm{Var}(X_N) = \sigma_X^2 = N\, p_X(1-p_X) + N(N-1)\Big(\prod_{r=1}^{m}\frac{(N-n_r)(N-n_r-1)}{N(N-1)} - \prod_{r=1}^{m}\Big(1-\frac{n_r}{N}\Big)^2\Big),$$

$$\mathrm{Var}(Y_N) = \sigma_Y^2 = N\, p_Y(1-p_Y) + N(N-1)\Big(\prod_{r=1}^{m}\frac{n_r\cdot(n_r-1)}{N\cdot(N-1)} - \prod_{r=1}^{m}\frac{n_r^2}{N^2}\Big),$$

$$\mathrm{Cov}(X_N, Y_N) = N\, p_Y + N(N-1)\, p_Y\Big(1 - \prod_{r=1}^{m}\frac{N-n_r}{N-1}\Big) - N^2 p_X p_Y.$$

Write

$$\mu_N = (\mu_X, \mu_Y), \qquad \Sigma_N = \begin{bmatrix} \sigma_X^2 & \mathrm{Cov}(X_N, Y_N) \\ \mathrm{Cov}(X_N, Y_N) & \sigma_Y^2 \end{bmatrix} \tag{4.5}$$

for the mean vector and covariance matrix of (X_N, Y_N).

Scaling and leading constants. Let $A := \prod_{r=1}^{m}(1-\alpha_r)$, $B := A^2$, $p_X := 1 - A$, and $p_Y := \prod_{r=1}^{m} \alpha_r$. With m and $\alpha_r \in (0,1)$ fixed,

$$\Sigma_N \;=\; N\,\Sigma(\alpha) + O(1), \qquad \Sigma(\alpha) = \begin{bmatrix} v_X(\alpha) & c_{XY}(\alpha) \\ c_{XY}(\alpha) & v_Y(\alpha) \end{bmatrix},$$

where

$$\begin{aligned} v_X(\alpha) &= p_X(1-p_X) \;-\; B\sum_{r=1}^{m} \frac{\alpha_r}{1-\alpha_r}, \\ v_Y(\alpha) &= p_Y(1-p_Y) \;-\; p_Y^2 \sum_{r=1}^{m} \frac{1-\alpha_r}{\alpha_r}, \\ c_{XY}(\alpha) &= -(m-1)\,A\,p_Y. \end{aligned}$$

Consequently, $\mathrm{Var}(X_N) = \Theta(N)$, $\mathrm{Var}(Y_N) = \Theta(N)$, $\mathrm{Cov}(X_N, Y_N) = \Theta(N)$, and $\Sigma_N/N \to \Sigma(\alpha)$ (positive definite for $\alpha_r \in (0,1)$).

Justification. Write $a_r := 1 - \alpha_r$. Using the uniform (in fixed α_r) expansions

$$\frac{(N-n_r)(N-n_r-1)}{N(N-1)} = a_r^2 + \frac{a_r^2 - a_r}{N-1}, \qquad \frac{n_r(n_r-1)}{N(N-1)} = \alpha_r^2 + \frac{\alpha_r^2 - \alpha_r}{N-1},$$

and

$$\prod_{r=1}^{m} \frac{N-n_r}{N-1} = A\Big(1 + \frac{m}{N-1}\Big) + O(N^{-2}),$$

a first-order product expansion in the variance/covariance formulas of §4.2 yields

$$\mathrm{Var}(X_N) = N\Big(p_X(1-p_X) - B\sum_{r=1}^{m} \frac{\alpha_r}{1-\alpha_r}\Big) + O(1),$$

$$\mathrm{Var}(Y_N) = N\Big(p_Y(1-p_Y) - p_Y^2 \sum_{r=1}^{m} \frac{1-\alpha_r}{\alpha_r}\Big) + O(1),$$

$$\mathrm{Cov}(X_N, Y_N) = N\Big(-(m-1)\,A\,p_Y\Big) + O(1),$$

i.e. $\Sigma_N = N\,\Sigma(\alpha) + O(1)$ and hence each entry is $\Theta(N)$ on any compact parameter set $\alpha_r \in [\varepsilon, 1-\varepsilon]$.

4.3 Bivariate CLT via Stein's method

Theorem 4.2. *Let X_N, Y_N be defined as in the model. Assume Σ_N is positive definite for all N, and is their exact finite-N covariance matrix. Define*

$$W_N := \Sigma_N^{-1/2} \begin{pmatrix} \frac{X_N - \mathbb{E}(X_N)}{\sqrt{N}} \\ \frac{Y_N - \mathbb{E}(Y_N)}{\sqrt{N}} \end{pmatrix}. \tag{4.6}$$

and let $Z \sim \mathcal{N}_2(0, I_2)$. Then

$$d_{\mathcal{C}}(W_N, Z) = O(N^{-1/2}), \tag{4.7}$$

where $d_{\mathcal{C}}$ denotes the distance induced by the test class $\mathcal{C}$ of indicators of multivariate convex sets.

Equivalently, if

$$G_N \sim \mathcal{N}_2(\mu_N, \Sigma_N) \tag{4.8}$$

then

$$d_{\mathcal{C}}\left(\frac{(X_N - \mathbb{E}(X_N), Y_N - \mathbb{E}(Y_N))}{\sqrt{N}}, \frac{G_N - (\mathbb{E}(X_N), \mathbb{E}(Y_N))}{\sqrt{N}}\right) = O(N^{-1/2}). \tag{4.9}$$

Note that

$$\frac{G_N - (\mathbb{E}(X_N), \mathbb{E}(Y_N))}{\sqrt{N}} = \mathcal{N}_2(0, \Sigma_N/N). \tag{4.10}$$

Proof. Using exchange pair reconstruction, write $V_k = (I_k - \mathbb{E}(I_k), J_k - \mathbb{E}(J_k))$ so that

$$(X_N - \mathbb{E}(X_N), Y_N - \mathbb{E}(Y_N)) = \sum_{k=1}^{N} V_k \tag{4.11}$$

Choose K uniformly from $\{1, 2, ..., N\}$ and resample the configuration for item K independently, yielding V'_K. Define

$$W'_N = W_N + \Delta, \qquad \Delta = \frac{1}{\sqrt{N}}\Sigma_N^{-1/2}(V'_K - V_K). \tag{4.12}$$

By construction, (W_N, W'_N) is exchangeable.

Conditioning on $V_1, ..., V_N$,

$$\mathbb{E}(\Delta | V_1, ..., V_N) = \frac{1}{\sqrt{N}}\Sigma_N^{-1/2}(\mu - V_K), \tag{4.13}$$

where $\mu = \mathbb{E}(V_k)$. Averaging over K gives

$$\mathbb{E}(\Delta | W_N) = -\frac{1}{N} W_N. \tag{4.14}$$

Thus $\lambda = 1/N$, $R = 0$.

Since each replacement changes at most one indicator vector, so $||V'_K - V_K|| \leq c$ for some constant c. Since $||\Sigma_N^{-1/2}||$ is uniformly bounded for large N,

$$|||\Delta|| \leq \frac{C}{\sqrt{N}}, \qquad \mathbb{E}(||\Delta||^3) = O(N^{-3/2}). \tag{4.15}$$

And $|\Delta_i \Delta_j| \leq \frac{C}{N}$ implies

$$\operatorname{Var}(\mathbb{E}(\Delta_i \Delta_j | W_N)) = O(N^{-3}) \tag{4.16}$$

and hence

$$\sqrt{\sum_{i,j} \mathrm{Var}(\mathbb{E}(\Delta_i \Delta_j | W_N))} = O(N^{-3/2}) \tag{4.17}$$

And applying multivariate Stein's theorem arrives:

$$d_{\mathcal{C}}(W_N, Z) \leq \frac{C}{N}[O(N^{-3/2} + O(N^{-3/2})] \leq \frac{C'}{\sqrt{N}} \tag{4.18}$$

for some constant C, C'. Hence $d_{\mathcal{C}}(W_N, Z) = O(N^{-1/2})$ □

Corollary 4.3. *Suppose additionally that*

$$\Sigma := \lim_{N \to \infty} \frac{\Sigma_N}{N} \tag{4.19}$$

exists and is positive definite. Then as $N \to \infty$,

$$\frac{(X_N - \mathbb{E}(X_N), Y_N - \mathbb{E}(Y_N))}{\sqrt{N}} \to \mathcal{N}_2(0, \Sigma), \tag{4.20}$$

or equivalently,

$$(X_N, Y_N) \to \mathcal{N}_2(\mu_N, N\Sigma) \tag{4.21}$$

Proof. From Theorem 4.2,

$$d_{\mathcal{C}}\left(\frac{(X_N - \mathbb{E}(X_N), Y_N - \mathbb{E}(Y_N))}{\sqrt{N}}, \mathcal{N}_2(0, \frac{\Sigma_N}{N})\right) \leq \frac{C}{\sqrt{N}} \tag{4.22}$$

for some constant C. Since $\Sigma_N/N \to \Sigma$ and the Gaussian law depends continuously on its covariance matrix in $d_{\mathcal{C}}$, the triangle inequality yields the claim. □

4.4 Univariate marginals

Corollary 4.4. *The univariate PMFs for the union and intersection are the marginals of the bivariate law:*

$$\mathrm{P}(X = u) = \sum_v p_{X,Y}(u, v), \qquad \mathrm{P}(Y = v) = \sum_u p_{X,Y}(u, v).$$

Proof. By Theorem 4.2, the univariate normal approximations for X and Y follow immediately by marginalisation of the bivariate Gaussian.

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2), \qquad Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2) \tag{4.23}$$

A self-contained combinatorial derivation of the univariate formulas (together with explicit feasibility ranges) is included in Appendix B for completeness. □

5 Jaccard index

Recall $X_N = \left|\cup_{r=1}^m P_r\right|$ and $Y_N = \left|\cap_{r=1}^m P_r\right|$. The Jaccard index is

$$J_N = \frac{Y_N}{X_N} \in [0,1] \qquad \big(\text{with } X_N > 0 \text{ in our model since } n_r > 0\big).$$

5.1 Exact finite-N law for $m = 2$

Proposition 5.1 (Exact pmf for $m = 2$)**.** *Let $m = 2$ and let*

$$v \in \big\{\max(0, n_1 + n_2 - N),\, \dots,\, \min(n_1, n_2)\big\}.$$

Then

$$\mathbb{P}\left(J_N = \frac{v}{n_1 + n_2 - v}\right) = \mathbb{P}(Y_N = v) = \frac{\binom{n_1}{v}\binom{N-n_1}{n_2-v}}{\binom{N}{n_2}}.$$

Consequently,

$$\mathbb{E}[J_N] = \sum_{v=\max(0,n_1+n_2-N)}^{\min(n_1,n_2)} \frac{v}{n_1+n_2-v}\cdot\frac{\binom{n_1}{v}\binom{N-n_1}{n_2-v}}{\binom{N}{n_2}} = \sum_v \frac{v}{n_1+n_2-v}\cdot\frac{F_2(n_1+n_2-v,\, v;\, S_2)}{\binom{N}{n_1}\binom{N}{n_2}},$$

where F_2 is given in (4.2)*.*

Proof. Condition on P_1; then $Y_N = |P_1 \cap P_2|$ is hypergeometric with parameters (N, n_1, n_2), yielding the stated pmf. The mapping $v \mapsto v/(n_1+n_2-v)$ is injective on the feasible integers, so

$$\mathbb{P}\left(J_N = \frac{v}{n_1+n_2-v}\right) = \mathbb{P}(X_N = n_1+n_2-v,\ Y_N = v).$$

The expectation follows by summation. □

5.2 Exact law for general m via the bivariate pmf

Proposition 5.2 (Exact pmf for general m)**.** *Fix coprime integers $a \geq 1$ and $0 \leq b \leq a$. Then*

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

where F_m is the exact count from Theorem 4.3 and

$$\mathcal{K}(a,b) = \Big\{ k \in \mathbb{Z}_{\geq 0} :\ (ka,\, kb) \in \mathcal{R} \text{ (the feasible region)} \Big\}.$$

Equivalently, one may bound k by any subset of the feasibility constraints, e.g.

$$k\,b \leq \min_{r\leq m} n_r, \qquad k\,a \leq N, \qquad \sum_{r=1}^m n_r \geq k\big(a+(m-1)b\big), \qquad \sum_{r=1}^m n_r \leq k\big((m-1)a+b\big).$$

Proof. The event $J_N = b/a$ is $\{Y_N/X_N = b/a\}$. Since X_N, Y_N are integers and $\gcd(a,b) = 1$, this holds iff there exists a (unique) $k \in \mathbb{Z}_{\geq 0}$ with $(X_N, Y_N) = (ka, kb)$. Hence

$$\mathbb{P}\Big(J_N = \frac{b}{a}\Big) = \sum_{k \geq 0} \mathbb{P}(X_N = ka,\ Y_N = kb).$$

The summands vanish unless (ka, kb) is feasible; substituting the bivariate pmf (4.1) yields the formula. The displayed bounds follow by inserting $(u, v) = (ka, kb)$ into the constraints defining $\mathcal{R}$. □

Remarks. (i) The support of J_N consists of rationals $b/a \in [0,1]$ in lowest terms for which some (ka, kb) is feasible; for $m = 2$ this reduces to the image of $v \mapsto v/(n_1 + n_2 - v)$. (ii) Mass at $J_N = 1$ occurs only when all sets are identical, which is feasible iff $n_1 = \cdots = n_m$; mass at $J_N = 0$ corresponds to $Y_N = 0$.

5.3 Gaussian approximation via the delta method

Recall $\mu_N = (\mu_X, \mu_Y)$ and Σ_N from §4.2, and set $g(x,y) = y/x$ on $\{x > 0\}$.

Theorem 5.3 (Delta–method Gaussian law for J_N)**.** *Assume m is fixed, $\alpha_r = n_r/N \in (0,1)$ are fixed, and $N \to \infty$ so that the bivariate CLT of Theorem 4.2 holds for (X_N, Y_N). Then*

$$d_W\Big(\frac{J_N - \mu_J}{\sigma_J},\, Z\Big) = O(N^{-1/2}), \quad Z \sim \mathcal{N}(0,1),$$

where d_W is the class of smooth test function for Wasserstein distance.

$$\begin{aligned}
\mu_J &= g(\mu_N) \;+\; \tfrac{1}{2}\operatorname{tr}\!\big(H_g(\mu_N)\,\Sigma_N\big) \;+\; O(N^{-1/2}) && (5.1)\\
&= \frac{\mu_Y}{\mu_X} \;+\; \frac{\mu_Y \operatorname{Var}(X_N) - \mu_X \operatorname{Cov}(X_N, Y_N)}{\mu_X^3} \;+\; O(N^{-1/2}), && (5.2)\\
\sigma_J^2 &= \nabla g(\mu_N)^\top \Sigma_N\, \nabla g(\mu_N) \;+\; O(N^{-1/2}) && (5.3)\\
&= \frac{\operatorname{Var}(Y_N)}{\mu_X^2} + \frac{\mu_Y^2}{\mu_X^4}\operatorname{Var}(X_N) - \frac{2\mu_Y}{\mu_X^3}\operatorname{Cov}(X_N, Y_N) + O(N^{-1/2}). && (5.4)
\end{aligned}$$

Here

$$\nabla g(\mu_N) = \Big(-\frac{\mu_Y}{\mu_X^2},\ \frac{1}{\mu_X}\Big), \qquad H_g(\mu_N) = \begin{bmatrix} \dfrac{2\mu_Y}{\mu_X^3} & -\dfrac{1}{\mu_X^2} \\ -\dfrac{1}{\mu_X^2} & 0 \end{bmatrix}.$$

Proof sketch. Theorem 4.2 gives $O(N^{-1/2})$ normal approximation for the standardised (X_N, Y_N). Applying the multivariate delta method (or Stein's method for smooth transforms) to $g(x,y) = y/x$ yields (5.2)–(5.4) and the $O(N^{-1/2})$ bound for J_N. □

Remarks. *Where the normal works best.* The sensitivity $\|\nabla g(\mu_N)\| \sim 1/\mu_X$ is smallest when μ_X is a healthy fraction of N (moderate overlaps), explaining the strong agreement seen in Figure 2a–b and Figure 3. Near the feasible boundaries ($J_N \approx 0$ or 1), curvature (entries of H_g) and lattice truncation make the approximation less accurate—consistent with the darker regions in Figure 2c. For $m = 2$ the exact finite–N distribution of J_N (Proposition 5.1) provides a benchmark; the delta–method mean in (5.2) tracks the exact $\mathbb{E}[J_N]$ closely across (n_1, n_2) (Figure 2). And when N grows, errors decreasing as N grows (Figure 3).

6 Numerical validation

All figures in this section use the *exact* finite–N counts from the paper: the bivariate recursion F_m in Theorem 4.3 for (X_N, Y_N) and its univariate specialisations C_m (unions) and D_m (intersections) from Appendix B. Gaussian overlays are parameterised by the closed-form moments from §4.2, and for Jaccard by the delta–method (μ_J, σ_J^2) in §5. Where a continuous normal density is drawn over a discrete pmf, we apply a half-cell continuity correction.

6.1 Bivariate distribution

Figure 1 shows the *exact* joint pmf $p_{X,Y}$ from (4.1) (via F_m) as a heat map, with isocontours of the Gaussian surrogate $\mathcal{N}_2(\mu_N, \Sigma_N)$ overlaid. The orientation and eccentricity of the ellipses are determined by the covariance $\mathrm{Cov}(X_N, Y_N)$ in §4.2; the close alignment of isocontours with pmf level sets empirically supports the approximation guaranteed by the bivariate CLT (Theorem 4.2), 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 §4.1), where lattice effects and truncation become non-negligible.

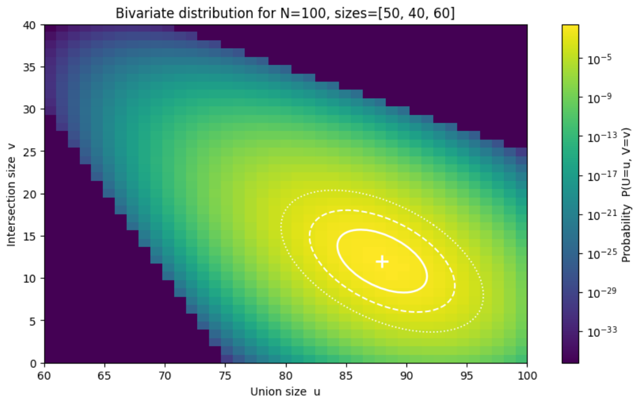
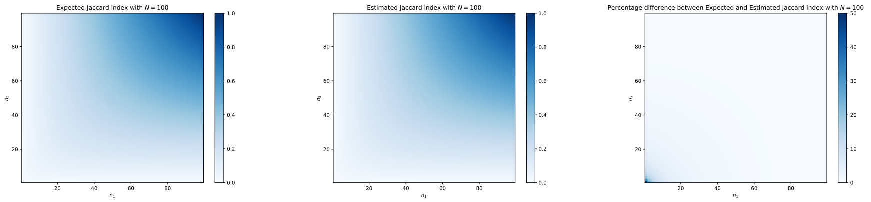


Figure 1: Exact joint pmf of (X_N, Y_N) (heat map via F_m) with $\mathcal{N}_2(\mu_N, \Sigma_N)$ isocontours (moments from §4.2) overlaid; continuity correction applied.

6.2 Jaccard distribution

We validate both the *exact* Jaccard constructions from §5 and the delta–method surrogate.

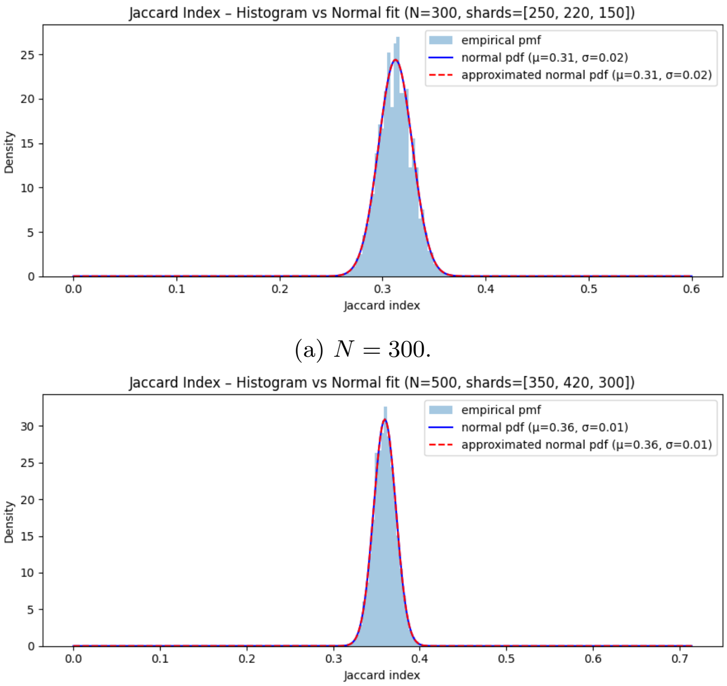
Expected value over (n_1, n_2)**.** Panel (a) of Figure 2 plots the exact $\mathbb{E}[J_N]$ for $m = 2$, obtained by summing the hypergeometric law of $V = |P_1 \cap P_2|$ (Proposition 5.1). Panel (b) shows the delta–method mean μ_J derived from (μ_N, Σ_N); panel (c) reports the relative error.



(a) Exact $\mathbb{E}[J_N]$ (Prop. 5.1) (b) Delta–method μ_J (§5) (c) $100 \times |\mathbb{E}[J_N] - \mu_J| / \mathbb{E}[J_N]$

Figure 2: Expected Jaccard index across (n_1, n_2) with $N = 100$. Exact (a) vs delta–method (b); relative error (c).

Empirical histogram vs two normal fits. Figure 3 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 §4.2.



(a) $N = 300$.

(b) $N = 500$.

Figure 3: Exact pmf of J_N (via Prop. 5.2) vs delta–method $\mathcal{N}(\mu_J, \sigma_J^2)$. Continuity correction and truncation at $[0, 1]$ applied to the overlay.

6.3 Univariate distribution

The univariate plots use the global recursions C_m and D_m from Appendix B, which are the marginals of F_m (cf. (4.1)).

Union. Figure 4 shows the exact pmf $\mathbb{P}(X_N = u)$ for $N = 10$ and $S_3 = (5, 6, 4)$. Mass at $u = N$ corresponds to full coverage (cf. the "collusion" event in §7); the mode sits close to μ_X from subsection 4.2.

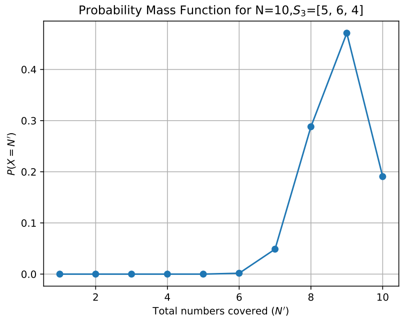


Figure 4: Exact pmf of X_N for $N = 10$ and $S_3 = (5, 6, 4)$ (via C_3).

For tail requirements in §7, we evaluate

$$\mathbb{P}(X_N \geq k) = \sum_{x=k}^{N} \frac{C_m(x; S_m)}{\prod_{n \in S_m} \binom{N}{n}}, \tag{6.1}$$

displayed in Figure 5 for the same parameters.

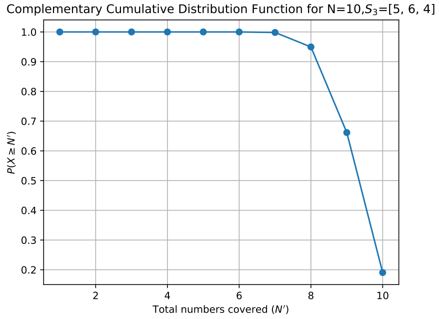
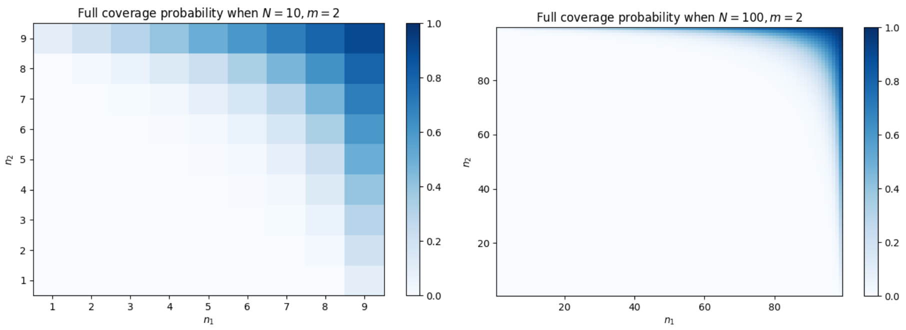


Figure 5: Tail probability $\mathbb{P}(X_N \geq k)$ for $N = 10$ and $S_3 = (5, 6, 4)$ (via C_3).

The full-coverage probability is

$$\mathbb{P}(X_N = N) = \frac{C_m(N; S_m)}{\prod_{n \in S_m} \binom{N}{n}},$$

and its dependence on (n_1, n_2) for $m = 2$ is summarised in Figure 6 (useful for merged-filter sizing in §7).



(a) $N = 10$

(b) $N = 100$

Figure 6: Full-coverage probability $\mathbb{P}(X_N = N)$ for $m = 2$ across (n_1, n_2) (via C_2).

Large-N behaviour. Theorem 4.2 gives an $O(N^{-1/2})$ quantitative CLT for (X_N, Y_N) after standardisation; univariate normal approximations follow by marginalisation. Figures 7 and 8 overlay $\mathcal{N}(\mu_X, \sigma_X^2)$ and $\mathcal{N}(\mu_Y, \sigma_Y^2)$ (moments from subsection 4.2) on the exact pmfs. Agreement is excellent away from the boundaries $x \approx N$ and $y \approx 0, \min_r n_r$; in extremely sparse-overlap regimes (very small p_Y) a Poisson or compound-Poisson surrogate for Y_N can be sharper, as noted in §7.

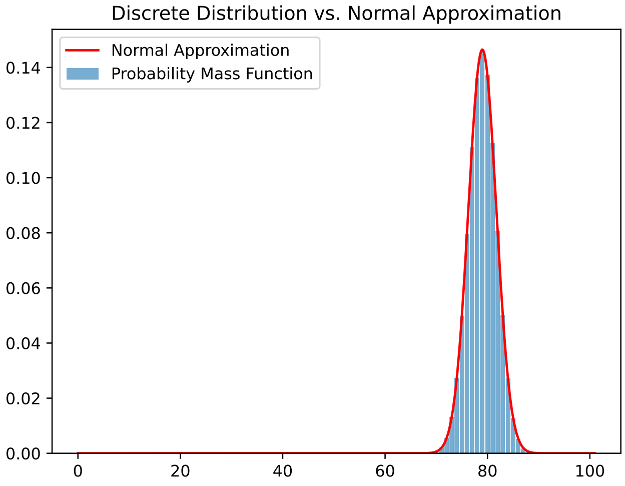


Figure 7: Exact pmf of X_N for $N = 100$ and $S_3 = (50, 30, 40)$ with normal overlay $\mathcal{N}(\mu_X, \sigma_X^2)$; continuity correction applied.

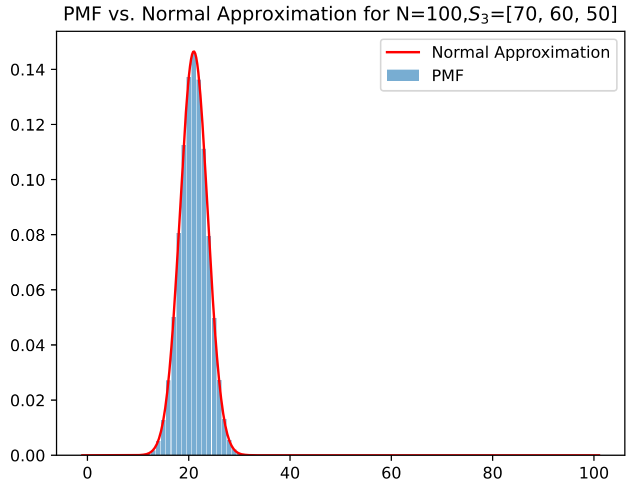


Figure 8: Exact pmf of Y_N for $N = 100$ and $S_3 = (70, 60, 50)$ with normal overlay $\mathcal{N}(\mu_Y, \sigma_Y^2)$; continuity correction applied.

7 Applications

This section shows how the exact counts F_m (and their univariate marginals), together with the moment and CLT results for X_N, Y_N, and J_N, translate into practical procedures. We write $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$ and Σ_N for the mean, variance, covariance of (X_N, Y_N) from §4.2. When exact probabilities are feasible we use F_m; otherwise we use the univariate/bivariate normal approximations with standard continuity corrections.

7.1 Probabilistic data structures

Probabilistic data structures (PDS) are compact, mergeable sketches that answer set-style queries with tunable error in sublinear space. Canonical examples include Bloom filters (approximate membership) and MinHash (Jaccard similarity). Our (X_N, Y_N) and J_N laws give direct, finite-N design rules.

7.1.1 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 X, is well-approximated by

$$\text{FPR}(X_) \;\approx\; \left(1 - e^{-hX/M}\right)^h.$$

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

$$\mathrm{P}\big(\mathrm{FPR}(X) \le \varepsilon\big) = \mathrm{P}\left(X \le \frac{M}{h} \log \frac{1}{1 - \varepsilon^{1/h}}\right) \;\ge\; 1 - \delta. \tag{7.1}$$

Two evaluation routes:

- *Exact:* compute the RHS by summing the union marginal of F_m: $\sum_{u \le x^\star} \sum_v \frac{F_m(u,v)}{\prod_r \binom{N}{n_r}}$, with $x^\star$ the threshold inside the braces.
- *Gaussian:* use $X \approx \mathcal{N}(\mu_X, \sigma_X^2)$ and $\mathrm{P}(X \le x^\star) \approx \Phi\big((x^\star + 0.5 - \mu_X)/\sigma_X\big)$.

For merge-heavy workloads, you can invert (7.1) to solve for the minimal M given h (or vice versa) at reliability $1 - \delta$.

7.1.2 MinHash: sample size planning with a J prior

A MinHash sketch [5] 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 (5.2) 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 [2,8], suppose a secret is split into N labeled shares; party r receives n_r distinct shares (uniformly, without replacement). Reconstruction depends on the *union* size $X = \big|\cup_r P_r\big|$ and, in some policies, on overlap levels.

All-or-nothing (N, N) additive sharing. Recovery occurs iff $X = N$. Evaluate

$$\mathrm{P}(X = N) = \sum_v \frac{F_m(N, v)}{\prod_r \binom{N}{n_r}} \quad \text{or} \quad \mathrm{P}(X = N) \approx 1 - \Phi\Big(\frac{N - 0.5 - \mu_X}{\sigma_X}\Big).$$

Design: choose $\{n_r\}$ (or N) to make this probability exceed a target reliability.

Threshold (k, N) schemes. Here recovery occurs if $X \geq k$. Compute

$$\mathrm{P}(X \geq k) = \sum_{u=k}^{N} \sum_{v} \frac{F_m(u,v)}{\prod_r \binom{N}{n_r}} \quad \text{or} \quad \mathrm{P}(X \geq k) \approx 1 - \Phi\Big(\frac{k - 0.5 - \mu_X}{\sigma_X}\Big).$$

If you also need a bound on *over-concentration*, impose $Y \leq L$ simultaneously and evaluate $\mathrm{P}(X \geq k,\ Y \leq L)$ either exactly by summing F_m over the rectangle, or via the bivariate normal CDF for (X_N, Y_N).

7.3 Incidence-graph operations

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

7.3.1 Joint SLAs: cover enough, avoid hotspots

Pick targets K and L and certify

$$\mathrm{P}(X \geq K,\ Y \leq L) \ \geq \ 1 - \delta.$$

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 $1 - \delta$.

7.3.2 Conditional redundancy from observed coverage

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

$$\mathbb{E}(Y \mid X = \widehat{x}) = \mu_Y + \rho\,\frac{\sigma_Y}{\sigma_X}(\widehat{x} - \mu_X), \qquad \mathrm{Var}(Y \mid X = \widehat{x}) = \sigma_Y^2(1 - \rho^2),$$

with $\rho = \mathrm{Cov}(X, Y)/(\sigma_X \sigma_Y)$. Use $\mathrm{P}(Y > L \mid X = \widehat{x})$ as an online alarm for redundancy spikes or collusion risk.

7.4 Implementation notes

- *Exact mode:* compute F_m via the recursion in §4.1 with memoisation and explicit bounds; obtain union/intersection marginals by summation over v or u.
- *Gaussian mode:* use μ_X, μ_Y, Σ_N from §4.2; apply continuity corrections when mapping to rectangles or half-lines.

8 Conclusion

We developed a unified distributional theory for coverage and full overlap in random set systems under fixed–cardinality, without–replacement sampling. Our first main contribution is an *exact* bivariate recursion $F_m(u, v; S_m)$ for the joint law of

$$(X_N, Y_N) = \Big(\Big|\bigcup_{r=1}^{m} P_r\Big|,\ \Big|\bigcap_{r=1}^{m} P_r\Big|\Big),$$

with explicit feasibility bounds that ensure valid, finite–N computation. The univariate laws for union and intersection appear as immediate *marginals*, eliminating the need for separate derivations. Second, using a swap–based exchangeable–pairs coupling, we proved a bivariate CLT for (X_N, Y_N) with $O(N^{-1/2})$ error, from which univariate normal approximations follow by marginalisation. Third, for the Jaccard index $J_N = Y_N/X_N$ we gave the exact finite–N distribution for $m = 2$, a construction for general m via F_m, and a delta–method Gaussian surrogate with explicit mean/variance in terms of (μ_X, μ_Y, Σ_N).

On the algorithmic side, a memoised implementation of F_m provides a practical evaluator for moderate sizes, and our experiments indicate that the Gaussian surrogates are accurate across a wide range of parameters, especially away from ultra–sparse corners. We illustrated how these tools translate into plug-and-play procedures for probabilistic data structures (Bloom sizing under merges; MinHash sample planning), incidence–graph design (joint coverage–overlap SLAs; conditional redundancy), and secret-sharing thresholds.

Limitations. Our analysis assumes (i) independent parties, (ii) fixed m, (iii) uniform sampling without replacement with prescribed $\{n_r\}$, and (iv) asymptotics in N. Exact evaluation of F_m is pseudo-polynomial in the size of the feasible grid and can become heavy for large m or extreme $\{n_r\}$. The CLTs target bulk accuracy; in very rare-event regimes (e.g., Y_N for large m) Poisson or compound-Poisson limits are sharper.

Overall, the paper provides exact finite–N formulas where they are tractable and principled Gaussian (or Poisson) surrogates where they are not, yielding a compact toolkit for both analysis and design in coverage/overlap problems. We hope the recursion F_m and the accompanying CLTs serve as a baseline for future refinements and broader models.

References

[1] Michael Barot and José Antonio de la Peña. Estimating the size of a union of random subsets of fixed cardinality. *Elemente der Mathematik*, 56(4):163–169, 12 2001.

[2] Amos Beimel. Secret-sharing schemes: A survey. In *Third international conference on Coding and cryptology*, volume 6639 LNCS, pages 11–46, 2011.

[3] Andrei Broder and Michael Mitzenmacher. Network applications of bloom filters: A survey. *Internet Mathematics*, 1(4):485–509, 2004.

[4] Sourav Chatterjee and Elizabeth Meckes. Multivariate normal approximation using exchangeable pairs. *Latin American Journal of Probability and Mathematical Statistics*, 4:237–283, 1 2008.

[5] Edith Cohen. Min-Hash Sketches. *Encyclopedia of Algorithms, Second Edition*, pages 1282–1287, 1 2016.

[6] Alex T. Kalinka. The probability of drawing intersections: extending the hypergeometric distribution. 5 2013.

[7] Gesine Reinert and Adrian Röllin. Multivariate normal approximation with Stein's method of exchangeable pairs under a general linearity condition. *Annals of Probability*, 37(6):2150–2173, 12 2009.

[8] Adi Shamir. How to share a secret. *Communications of the ACM*, 22(11):612–613, 11 1979.

A Appendix: Moment derivations

A.1 Set union

Define indicator variables on the universe $[N]$. For each $i \in \{1, \dots, N\}$ let

$$I_i = \begin{cases} 1, & \text{if element } i \text{ is selected by at least one party,} \\ 0, & \text{otherwise.} \end{cases}$$

Then the union size is $X = \sum_{i=1}^{N} I_i$. For a fixed element i and party r, $\mathbb{P}(i \in P_r) = n_r/N =: \alpha_r$, independently across r. Hence

$$\mathbb{E}(X) = \sum_{i=1}^{N} \mathbb{E}(I_i) = \sum_{i=1}^{N} \mathbb{P}(I_i = 1) = N\Big(1 - \prod_{r=1}^{m}(1 - \alpha_r)\Big) =: N\, p_X.$$

Variance and normal approximation. To assess variability,

$$\mathrm{Var}(X) = \sum_{i=1}^{N} \mathrm{Var}(I_i) + \sum_{i \neq j} \mathrm{Cov}(I_i, I_j) = N\, p_X(1 - p_X) + N(N-1)\ \mathrm{Cov}(I_1, I_2),$$

by symmetry. For distinct $i \neq j$,

$$\begin{aligned} \mathbb{E}[I_i I_j] &= \mathbb{P}(I_i = 1,\ I_j = 1) = 1 - 2\,\mathbb{P}(I_i = 0) + \mathbb{P}(I_i = 0,\ I_j = 0) \\ &= 1 - 2\Big(\prod_{r=1}^{m}(1 - \alpha_r)\Big) + \prod_{r=1}^{m} \frac{\binom{N-2}{n_r}}{\binom{N}{n_r}}, \end{aligned}$$

so

$$\begin{aligned}\mathrm{Cov}(I_i, I_j) &= \mathbb{E}[I_i I_j] - \mathbb{E}[I_i]\mathbb{E}[I_j] \\ &= -\Big(1 - p_X\Big)^2 + \prod_{r=1}^{m} \frac{\binom{N-2}{n_r}}{\binom{N}{n_r}} = -\prod_{r=1}^{m}(1-\alpha_r)^2 + \prod_{r=1}^{m} \frac{(N-n_r)(N-n_r-1)}{N(N-1)}.\end{aligned}$$

Using

$$\frac{(N-n_r)(N-n_r-1)}{N(N-1)} = (1-\alpha_r)^2 + \frac{(1-\alpha_r)^2 - (1-\alpha_r)}{N-1} = (1-\alpha_r)^2 + O(N^{-1}),$$

and fixed m, we get $\mathrm{Cov}(I_i, I_j) = \mathcal{O}(N^{-1})$. Therefore

$$\begin{aligned}\mathrm{Var}(X) &= N\,p_X(1-p_X) \;+\; N(N-1)\left(-\prod_{r=1}^{m}(1-\alpha_r)^2 + \prod_{r=1}^{m} \frac{(N-n_r)(N-n_r-1)}{N(N-1)}\right) \\ &= \Theta(N), \qquad\qquad (A.1)\end{aligned}$$

provided m is fixed and each α_r stays in $(\varepsilon, 1-\varepsilon)$.

A.2 Set intersection

For each $i \in \{1, \dots, N\}$ let

$$J_i = \begin{cases} 1, & \text{if element } i \text{ is selected by all parties,} \\ 0, & \text{otherwise.} \end{cases}$$

Then the intersection size is $Y = \sum_{i=1}^{N} J_i$. For a fixed element i, independence across parties gives

$$p_Y := \mathbb{P}(J_i = 1) = \prod_{r=1}^{m} \frac{n_r}{N} = \prod_{r=1}^{m} \alpha_r, \qquad \mathbb{E}(Y) = N\,p_Y.$$

Variance and normal approximation. Similarly,

$$\mathrm{Var}(Y) = N\,p_Y(1-p_Y) + N(N-1)\ \mathrm{Cov}(J_1, J_2).$$

For distinct $i \neq j$,

$$\mathbb{E}[J_i J_j] = \mathbb{P}(J_i = 1,\ J_j = 1) = \prod_{r=1}^{m} \frac{\binom{N-2}{n_r-2}}{\binom{N}{n_r}} = \prod_{r=1}^{m} \frac{n_r(n_r-1)}{N(N-1)},$$

hence

$$\mathrm{Cov}(J_i, J_j) = \prod_{r=1}^{m} \frac{n_r(n_r-1)}{N(N-1)} - \prod_{r=1}^{m} \Big(\frac{n_r}{N}\Big)^2 = \prod_{r=1}^{m} \left(\alpha_r^2 + \frac{\alpha_r^2 - \alpha_r}{N-1}\right) - \prod_{r=1}^{m} \alpha_r^2 = O(N^{-1}).$$

Thus

$$\begin{aligned}\mathrm{Var}(Y) &= N\,p_Y(1-p_Y) + N(N-1)\left(\prod_{r=1}^{m} \frac{n_r(n_r-1)}{N(N-1)} - \prod_{r=1}^{m} \Big(\frac{n_r}{N}\Big)^2\right) \\ &= \Theta(N) \quad (\text{for fixed } m \text{ and } \alpha_r \in (\varepsilon, 1-\varepsilon)). \qquad (A.2)\end{aligned}$$

B Appendix: Univariate

Throughout, tuples $(P_1, \dots, P_m)$ are *ordered* and all binomial coefficients obey the zero convention $\binom{a}{b} = 0$ when $b \notin \{0, 1, \dots, a\}$ (or $a < 0$). Write $S_m = (n_1, \dots, n_m)$.

B.1 Set union

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

$$\mathbb{P}(X = x) = \mathbb{P}(|P_1 \cap P_2| = v) = \frac{\binom{n_1}{v}\binom{N-n_1}{n_2-v}}{\binom{N}{n_2}}, \qquad v \in \{\max(0, n_1+n_2-N), \dots, \min(n_1, n_2)\}.$$

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

$$C_2(x; S_2) = \binom{N}{n_1}\binom{n_1}{v}\binom{N-n_1}{n_2-v}, \quad \text{so} \quad \mathbb{P}(X = x) = \frac{C_2(x; S_2)}{\binom{N}{n_1}\binom{N}{n_2}}.$$

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

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

$$C_3(x; S_3) = \sum_u \binom{u}{n_3 + u - x}\binom{N-u}{x-u} C_2(u; S_2),$$

hence

$$\mathbb{P}(X = x) = \frac{C_3(x; S_3)}{\prod_{i=1}^{3}\binom{N}{n_i}}.$$

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

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

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

$$C_1(x; S_1) = \begin{cases} \binom{N}{n_1}, & x = n_1, \\ 0, & \text{otherwise,} \end{cases}$$

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

$$C_m(x; S_m) = \sum_u \binom{u}{n_m + u - x}\binom{N-u}{x-u} C_{m-1}(u; S_{m-1}). \tag{B.1}$$

A convenient feasible range is

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

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

$$\mathbb{P}(X = x) = \frac{C_m(x; S_m)}{\prod_{i=1}^{m} \binom{N}{n_i}}. \tag{B.2}$$

B.2 Set intersection

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

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

Equivalently, the global count is

$$D_2(y; S_2) = \binom{N}{n_1}\binom{n_1}{y}\binom{N-n_1}{n_2-y}, \quad \text{so} \quad \mathbb{P}(Y = y) = \frac{D_2(y; S_2)}{\binom{N}{n_1}\binom{N}{n_2}}.$$

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

$$D_3(y; S_3) = \sum_{v_2} \binom{v_2}{y}\binom{N-v_2}{n_3-y} D_2(v_2; S_2), \qquad \mathbb{P}(Y = y) = \frac{D_3(y; S_3)}{\prod_{i=1}^{3} \binom{N}{n_i}}.$$

A convenient explicit range is

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

(Outside this range the binomials vanish.)

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

$$D_1(y; S_1) = \begin{cases} \binom{N}{n_1}, & y = n_1, \\ 0, & \text{otherwise}, \end{cases}$$

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

$$D_m(y; S_m) = \sum_{v_{m-1}} \binom{v_{m-1}}{y}\binom{N - v_{m-1}}{n_m - y} D_{m-1}(v_{m-1}; S_{m-1}). \tag{B.3}$$

A convenient feasible range is

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

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

$$\mathbb{P}(Y = y) = \frac{D_m(y; S_m)}{\prod_{i=1}^{m} \binom{N}{n_i}}. \tag{B.4}$$

Support. From standard feasibility constraints,

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