Soft Computing Fusion with Applications

www.scfa.reapress.com

Soft. Comput. Fusion. Appl. Vol. 3, No. 2 (2026) 99–113.

Paper Type: Original Article

Communication-Efficient Federated Rare-Pattern Mining under Non-IID Data and Privacy Constraints

Jyotsana Grover[1,*], Shakila Devi G T[2]

[1] Department of Computer Science and Information Systems, Birla Institute of Technology and Science, Pilani, India; jyotsana.grover@pilani.bits-pilani.ac.in.

[2] Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, India; shakiladevigt.sse@saveetha.com.

Citation:

Received: 10 September 2025 *Revised: 19 January 2026* *Accepted: 21 April 2026*	Grover, J., & Devi G T, Sh. (2026). Communication-efficient federated rare-pattern mining under non-IID data and privacy constraints. *Soft Computing Fusion with Applications, 3*(2), 99-113.

Abstract

Rare co-occurrence patterns can be difficult to discover across institutions whose transactions remain distributed. We present FedRPM, a bounded-candidate protocol combining noisy summaries, pooled support estimation, and a verification stage. The reported proof-of-concept campaign uses twenty eight-client federations of UCI Adult under two Dirichlet skew levels, ten seeds per level, six planted carrier-concentrated itemsets per run, and a centralized Apriori support-band reference. FedRPM achieves mean recall 0.9564 and precision 0.9999 for supports between 0.5 and 2.0% at 67.22 KB of serialized communication, compared with recall 0.9234, precision 0.9896, and 72.49 KB for naive broadcast (reported paired Wilcoxon W = 6, p = 2.67e-05). The full-lattice federated reference reaches recall 0.9926 at 82.59 KB, or 22.9% more communication than FedRPM. Reducing the nominal per-coordinate epsilon from 8 to 0.5 lowers recall to 0.9221. These epsilon settings describe count-noise scales, not a demonstrated total record-level privacy budget. Secure aggregation is simulated, transmitted-content bytes are an exposure surrogate rather than a privacy guarantee, and adaptive thresholds and the feasibility controller are untested. The results therefore characterize candidate coverage and communication in a small simulation; an audited transcript accountant and the experiment artifacts are required to establish end-to-end privacy and reproducibility.

Keywords: Federated pattern mining, Rare itemsets, Non-IID data, Differential privacy, Secure aggregation, Communication efficiency.

1 | Introduction

Rare co-occurrence patterns combine items, events, or attributes observed in a small fraction of transactions. They may support hypothesis generation in pharmacovigilance or fraud analytics when signals are distributed across institutions. Association analysis formalizes itemsets, supports, and confidence-thresholded rules [1], and rare-itemset research studies the low-support regime [2]. A globally rare pattern may be moderately

Corresponding Author: jyotsana.grover@pilani.bits-pilani.ac.in

https://doi.org/10.22105/scfa.v3i2.103

Licensee System Analytics. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0).

frequent at a few clients or thinly distributed. Legal, contractual, and organisational conditions can constrain pooling, motivating distributed discovery when an authorised centralized analysis is unavailable.

The algorithmic and systems ingredients for an alternative exist separately. Frequent-pattern mining has matured from level-wise candidate generation in Apriori [3] to compressed, prefix-tree counting in FP-growth [4], and rare-itemset methods adapt these foundations to bounded-support bands [2]. In parallel, federated learning has established that decentralised statistics can approximate pooled model training without moving raw data [5], with dedicated work on communication efficiency [6] and secure aggregation [7], and with comprehensive surveys of open problems [8], [9]. These developments, however, target parameter estimation rather than enumeration-style discovery, in which the output space is the power set of the item universe and the goal is to certify which patterns satisfy a support band. Federated pattern discovery therefore cannot simply reuse federated averaging: the objects exchanged, the pruning rules, and the error semantics all change.

Three obstacles motivate the design. Non-IID clients have unequal local supports, requiring record-count-weighted aggregation rather than an unweighted average [10]. Privacy noise interacts with low support [11], while candidate exchange grows with the combinatorial itemset universe. FedRPM investigates bounded candidate coverage, noisy support estimates, and communication in one protocol. The reviewed methods provide its components; the present evidence does not establish that every prior protocol lacks this combination or that the executed variant satisfies a total record-level differential-privacy guarantee.

The contribution combines a protocol specification with a reported simulation campaign. Its formal privacy design, candidate-selection heuristic, and measured realization are distinguished throughout.

A formal problem statement for federated rare-band pattern discovery that makes explicit the three-way coupling between rarity, privacy noise, and communication, expressed as a support-band semantics over non-IID client databases.

FedRPM uses bounded candidate screening, pooled noisy counts, and focused verification. A private implementation additionally requires calibrated releases for every data-dependent candidate and all observable metadata.

A design-level signal-to-noise inequality relates pool size, the count-stage budget, and federation size. It guides follow-up query design but does not certify support-band membership.

A proposed skew-adaptive threshold links reporting to privatized marginal distance; its direction, bias and communication benefit require evaluation.

The reported UCI Adult campaign compares five methods on twenty simulated federations, with ablations, a per-coordinate noise sweep, and paired exploratory statistical summaries.

The remainder of the manuscript is organised as follows. Section 2 reviews related work and synthesises the gap. Section 3 states research questions, hypotheses, the system model, and the formal problem. Section 4 presents the FedRPM methodology, its privacy accounting, complexity, and failure conditions. Section 5 reports the executed experimental campaign and its analysis, Section 6 discusses limitations and counterarguments, and Section 7 concludes with empirically grounded findings and a concrete roadmap.

2 | Related Work

2.1 | From Frequent to Rare: Centralized Pattern Discovery

Apriori exploits the anti-monotonicity of a minimum-support constraint [3], and FP-growth replaces repeated scans with prefix-tree pattern growth [4]. Emerging patterns identify support contrasts between populations [12]. Rare-band mining adds an upper support ceiling [2]: Once an itemset satisfies this ceiling, every superset also satisfies it, whereas the lower support condition is anti-monotone. These distinct monotonicities must be respected by candidate pruning.

The lower support edge permits Apriori pruning; the upper edge is applied when classifying complete candidate counts. Candidate truncation still restricts recall. Centralized algorithms can provide a pooled reference when pooling is authorised, while distributed methods change the information available to the coordinator.

2.2 | Federated Learning under Statistical Heterogeneity

Federated learning established coordinator-based training and communication-efficient updates [5], [6]; heterogeneous-data studies characterize local/global model discrepancies [10], [13]. For support estimation, the relevant distinction is simpler: the global support is the sum of local counts divided by the sum of local record counts. Non-IID data do not bias this exact weighted sum, but missing contributions or altered weighting can.

Secure aggregation can hide individual vectors while revealing their sum under a specified cryptographic threat model [7]. Count-min and count sketches provide bounded-size approximate summaries [14]. These ingredients motivate candidate screening and pooled counting, but a sketch alone does not calibrate record-level privacy or guarantee rare-pattern coverage. Published federated IoT infrastructure studies also consider blockchain and offloading [15]; their architectural security objectives are distinct from transcript-wide record-level privacy accounting.

2.3 | Privacy-Preserving Pattern Mining and Distributed Perturbation

Privacy-preserving data mining predates differential privacy. Early perturbation approaches reconstructed aggregate distributions from transformed values [16], while the cryptographic line used secure multi-party computation, beginning with two-party distributed mining [17], to mine association rules over vertically [18] and horizontally partitioned data [19]; contemporary surveys catalogued these models and their assumptions [20]. These protocols typically target exact results and pay for exactness with round complexity and strong collusion assumptions, while providing no calibrated guarantee against inference from the outputs themselves. Differential privacy supplies exactly such a calibration [11], and its adoption in machine learning, most prominently differentially private stochastic gradient descent [21], established accounting machinery for iterative releases.

In the local model, randomised-response systems such as RAPPOR demonstrated practical frequency estimation under per-record perturbation [22], and lower-bound results quantified the inherent cost of local privacy for estimation accuracy [23]. Closest to our setting, locally differentially private frequent-itemset mining protocols have been proposed for single-party collections of users [24], and heavy-hitter identification under local privacy has received substantial theoretical attention. These works, however, address frequent or heavy patterns at a single curator; the rare band we target lies far below the heavy-hitter regime, and federation adds a second layer of skew, since global rarity is compatible with local frequency. Moreover, iterative discovery requires composition-aware accounting across rounds, for which the composition literature provides the tools [11], [25].

2.4 | Gap Synthesis

Table 1 compares the reviewed approach families. FedRPM combines candidate truncation, noisy counting, and verification for a rare support band. The table describes these design choices and the stated assumptions; the executed campaign supplies no cryptographic implementation or full-transcript privacy audit.

Table 1. Comparison of approach families relevant to federated rare-pattern mining.

Approach Family	Representative Works	Data Placement	Privacy Regime	Rare-Band Support
Centralized pattern mining	[2-4], [12]	Pooled	None (trust the curator)	Exact when pooling is authorised.
Cryptographic multi-party mining	[17–19]	Partitioned	Cryptographic, no output guarantee	Exact at high communication cost
Local-model DP mining	[22], [24]	Individual randomized responses	Per-record local DP	Targets heavy patterns; rare band out of scope
Federated learning (models)	[5-7], [13]	Partitioned	Optional central DP	No enumeration semantics
FedRPM (this work)	Sections 3 and 4	Partitioned, non-IID	Conditional composed record DP; simulation unverified	Rare band with feasibility control

3 | Problem Formulation and Research Questions

3.1 | Research Questions and Hypotheses

The first research question concerns identifiability under noise: RQ1, under non-IID partitions of a transaction population across clients, can itemsets in a prescribed low-support band be identified with recall close to a centralized oracle while per-round communication is bounded by compact summaries rather than enumerated candidate universes? This question separates the statistical problem (support estimation at rare scales) from the systems problem (bytes exchanged per round), and our protocol couples them through the feasibility condition derived in Section 4.3. Concretely, the question is whether the recall-communication frontier of the proposed protocol dominates those of the baselines in Section 5, and how that dominance shifts as partition skew grows.

The second research question concerns the privacy interaction: RQ2: How does the privacy regime trade off against rare-band recall at a fixed communication budget, and where is the noise floor below which band membership becomes statistically unresolvable? We distinguish a secure-aggregation-only regime, where the server learns aggregate supports but no per-client values, from a full per-record differential-privacy regime where even aggregates are perturbed. The hypothesis is that the first regime incurs negligible recall loss relative to a no-privacy federation, whereas the second exhibits a degradation that sharpens as the band's lower edge approaches the noise floor.

The third research question concerns threshold adaptation: RQ3: How should per-client reporting thresholds adapt to local distributional skew so that communication is spent where it is informative, without losing patterns whose support is concentrated on a minority of clients? The competing failure modes are flooding (thresholds too low, so that clients report noise-dominated counts) and starvation (thresholds too high, so that thinly supported patterns are dropped before aggregation). We study these questions through three hypotheses.

H1 asks whether bounded-candidate screening and refinement provide a favourable recall-communication trade-off under the specified Dirichlet partitions. Asymptotic sublinear communication remains a separate scaling hypothesis.

H2 asks how properly composed record-level noise changes detection and support-band decisions relative to a secure-aggregation-only arm. The present per-coordinate sweep examines only the noise-scale part of this question.

H3 asks whether skew-adaptive thresholds improve the recall-communication frontier over fixed thresholds. The executed fixed-threshold campaign does not test this hypothesis.

3.2 | System Model and Notation

There are m clients with databases D_i over a shared item universe, n_i records per client, and total N. The privacy design uses add/remove-one-record adjacency within one client, treating public federation-size information as fixed auxiliary information; any data-dependent disclosure of sizes requires separate protection. Each exact itemset count has sensitivity one, but a vector of K counts can have L1 sensitivity K. Clients are institutional curators, not individual-record local-DP respondents. The total client ledger E must reserve budget for candidate discovery, histograms, counts, verification, and observable participation. Reliable delivery, no round dropouts, and aligned item identifiers are assumed.

Table 2. Notation used throughout the manuscript.

Symbol	Meaning	Domain/Unit
m, i	Number of clients, client index	Integer
D_i, n_i, N	client database, its size, federation size	Transactions
I, X	item universe, itemset	2^I
c_i(X)	Local count of X at client i	Records
s(X)	Global relative support of X	[0, 1]
theta_min, theta_rare	Band lower edge and rareness ceiling	[0, 1]
gamma	Minimum rule confidence for derived rules	[0, 1]
E, E_count, eps	E: whole transcript; E_count: count stage; eps: one query	Privacy units
Q_t, K_t	Candidate pool at round t and its cardinality	Itemsets
z_i(X)	Noisy count released by client i for X	Records
tau_i(t)	Per-client reporting threshold at round t	Records
d_i(t)	Client i's distance from the federated mixture	[0, 1]
V_t	Communication volume in round t	Bytes
R	Number of protocol rounds	Integer

3.3 | Formal Problem Statement

Global relative support is the record-weighted sum in *Eq. (1)*. A coordinator can obtain this sum by secure aggregation without seeing each raw client count. Approximation arises from privacy noise, sketch error, candidate truncation, or incomplete contribution, rather than from federation itself.

$$s(X) = \frac{\Sigma_i\, c_i(X)}{N}, \qquad c_i(X) = |\{T \in D_i : X \subseteq T\}|. \tag{1}$$

The evaluated target is a support-band itemset. Association-rule confidence is an optional downstream criterion: For a nonempty antecedent X with s(X)>0 and a disjoint consequent item y, conf(X -> y)=s(X union {y})/s(X). No confidence requirement is used to construct the experimental band.

$$X \in \text{Rare} \Leftrightarrow \theta_{\min} \leq s(X) \leq \theta_{\text{rare}}. \tag{2}$$

The design seeks a small symmetric difference between the returned and true support-band sets under communication and computational budgets. A private deployment additionally requires a composed record-level guarantee for the complete observable transcript. Estimated band decisions are not certificates unless error bounds, candidate coverage, and decision margins are supplied. The campaign evaluates the support-band and communication objectives, rather than this complete constrained problem.

4 | Proposed Methodology: The FedRPM Protocol

4.1 | Architecture Overview

The design separates candidate screening, pooled estimation, and focused refinement. Each data-dependent release must enter the transcript accountant. Section 5 specifies a sampled-count realization that differs from the digest and adaptive-threshold design below.

Stage A proposes a bounded candidate pool from private summaries and a public analyst vocabulary. Hashing a data-dependent candidate list does not make it private; a calibrated discovery mechanism and its budget are required.

Stage B pools noisy candidate counts. Cryptographic secure aggregation hides client vectors under its threat assumptions; the simulation models only the aggregate interface.

Stage C re-queries near-edge candidates using a reserved verification budget. A smaller pool can reduce per-coordinate noise under an equal split of a fixed stage budget, but does not itself validate a classification.

4.2 | Candidate Pooling and Sketched Screening

Locally concentrated patterns motivate candidate pooling. A curator may privately propose candidates using a calibrated mechanism; randomized-response frequency estimation [22], [24] cannot be assumed to privatize an arbitrary data-dependent heavy-itemset digest. Candidate discovery, digest length, and identifiers must have an explicit release model and reserved budget.

Diffuse patterns may enter through a public vocabulary or expansion of privately released summaries. Candidate truncation can miss patterns. Sketch collisions and count approximation add error distinct from privacy noise [14]; the executed sampled-coordinate summary has no count-sketch collision mechanism and does not test width/depth bounds.

4.3 | Privacy Layer and Detection Feasibility

For exact counts under the stated adjacency, allocate a count-stage budget E_count separately from candidate discovery and verification. Equal splitting across R count rounds and K_t coordinates gives *Eq. (3)*; adaptive query choices based only on the preceding private transcript can compose conditionally [11], [25]. Secure aggregation provides confidentiality under its own assumptions [7], not a substitute for this accountant. Stage budgets must satisfy E_discovery+E_count+E_verify+E_other≤E.

$$z_i(X) = c_i(X) + \eta_i(X), \qquad \eta_i(X) \sim \text{Laplace}\left(\frac{K_t R}{E_{\text{count}}}\right). \tag{3}$$

For independent equal-scale client noise, the support-noise standard deviation is sigma_t=(K_t R/E_count)sqrt(2m)/N. *Eq. (4)* is an approximate signal-to-noise screen using a Gaussian quantile, not a necessary detectability boundary or a confidence certificate. Reliable band membership depends on the distance to both band edges, candidate-selection error, sketch error, and any skipping bias, with simultaneous error control if many candidates are classified.

$$\theta_{\min} \geq z_{1-\alpha} \frac{K_t R}{E_{\text{count}}} \frac{\sqrt{2m}}{N}. \tag{4}$$

The illustrative *Algorithm 1* makes submission conditional on a previously privatized count. A new candidate has no previous count and must be queried without this skip gate. Gating on an earlier noisy value can be post-processing only when that value is covered by the ledger; it does not bound the skipped true count by the threshold. A reduction in K_t reduces standard deviation linearly and variance quadratically under *Eq. (3)*.

Algorithm 1. FedRPM communication round t.

Input: pool Q_t, count-stage budget E_count, reserved R rounds, band edges, thresholds tau_i(t), noise margin delta_t
1: server broadcasts Q_t and hash seeds to all clients
2: for each client i = 1..m in parallel do
3: for each candidate X in Q_t do
4: c_i(X) <- exact local count over D_i; sketches need a separate sensitivity bound
5: z_i(X) <- c_i(X)+Laplace(K_t R/E_count) // exact count sensitivity one
6: a_i(X) <- z_i(X) if X is new or prior private z_i(X)>=tau_i(t); else 0
7: end for
8: end for
9: secure aggregation returns S(X)=sum_i a_i(X), including gated zeros
10: s_hat(X)<-S(X)/N; report provisional support-band classification
11: for X with s_hat(X) within delta_t of either band edge do
12: queue focused re-query of X into Q_{t+1} with reduced pool budget
13: end for
14: update pool: drop above-band and resolved candidates; admit new
Output: rare-band estimate and refined pool Q_{t+1}

4.4 | Skew-Adaptive Thresholds and Tiered Refinement

A skew-adaptive threshold can be formed from privately released, normalized item-marginal distributions and their total-variation distance d_i(t). Their release costs must be charged before the threshold is treated as post-processing. *Eq. (5)* is a proposed heuristic whose performance and bias require evaluation.

$$\tau_i(t) = \left\lceil \theta_{min}\, n_i \left(1 + \kappa\, d_i(t)\right) \right\rceil. \quad (5)$$

The proposed heuristic raises thresholds for clients distant from the aggregate mixture. Marginal distance does not establish where itemset support is concentrated or diffuse, and the direction of this rule is not justified by a theorem. Because a noisy prior count may underestimate any true count, zero-submitted skips have no deterministic true-count bias bound of tau_i(t). Even a deterministic skip bound would accumulate across omitted clients. The executed threshold tau=1 differs from *Eq. (5)*, and neither the adaptive rule nor a bias controller was tested.

4.5 | Complexity, Assumptions, and Failure Conditions

Direct exact counting of K_t arbitrary candidates of bounded length ell takes O(n_i K_t ell) work, although a suitable trie, bitmap index, or sketch may reduce practical cost. Counter payload is O(K_t b) per client, plus candidate identifiers, discovery messages, verification, and cryptographic overhead. Server summation costs O(m K_t), while candidate generation and ranking add work. An (E,0) record guarantee is conditional on calibration and composition of every release, including discovery and metadata, and is not established for the executed variant. Independence of records is a modelling assumption for statistical interpretation, not a requirement of differential privacy.

The protocol has identifiable failure conditions, which the evaluation protocol is designed to expose rather than hide. Each condition is tied to a specific mechanism in Sections 4.2-4.4, so that a negative result in the planned campaign can be attributed to a component rather than to the design as a whole. The conditions most likely to bind in practice are the following:

I. A failed signal-to-noise screen indicates a difficult design point; it does not prove that every band decision is statistically unresolvable.

II. Screening miss: Diffuse rare patterns absent from every client's heavy digests and from the analyst vocabulary are never pooled and are silently lost; recall ceilings in the evaluation quantify this coverage gap.

III. Sketch approximation or inadequate sampled-coordinate coverage can impair candidate ranking; the executed pair-sampling sweep evaluates coverage, not count-sketch depth.

IV. Adversarial clients: a client inflating its counts can push itemsets into the band; robustness to such behaviour is out of scope here and flagged as future work.

5 | Experimental Results and Analysis

5.1 | Setup: Corpus, Federation, Methods, and Protocol

The campaign uses UCI Adult [26]. The source reports 45,222 complete-case records discretized into 23 binary items, followed by 8,000-basket seeded subsamples partitioned among eight clients using Dirichlet-style marginal regions [27]. Two skew levels, 0.5 and 1.0, with ten seeds per level are reported. Six carrier-concentrated itemsets are planted at target supports of 0.8, 1.0, 1.2, 1.5, 1.7, and 1.9%. Raw preprocessing, carrier allocation, and realized-support checks require the original script; planting modifies co-occurrences and may produce additional derived patterns.

Ground truth is computed by a compact bitmask Apriori implementation [3] on the pooled 8,000 baskets at minimum support equal to the ceiling of 0.005 times n (40 baskets); the rare band is the set of 2- and 3-itemsets with global support in [0.5%, 2.0%]. Per run the band contains on average 545.5 itemsets (19.8 pairs and 525.7 triples), of which a floor-adjacent subset of 108.9 itemsets lies in [0.5%, 0.7%], the hardest stratum of the band. Five methods are compared on every federation. The centralized Apriori oracle (the pooled upper bound; its communication is the raw pickled transaction matrix and its leakage equals that upload) realizes the centralized baselines of the design protocol - the FP-growth variant is algorithmically equivalent for band enumeration and was not reimplemented. Local-only mining runs per-client Apriori at a 1% local minimum support and unions the reported patterns; naive broadcast runs per-client Apriori at a 0.5% local minimum support, uploads every locally frequent itemset with its exact count, and lets the server sum the counts and apply the band filter. The full-lattice federated reference (communication-unconstrained) sketches all 253 pairs and all 1,771 triples with noisy secure-aggregated counts and applies the same verification stage as FedRPM. The sampled round-robin baseline of the design protocol was not executed.

FedRPM is executed in a four-round form. Round 1 sketches the 23 singletons plus 160 of the 253 item pairs with Laplace-noised, secure-aggregated counts at epsilon = 8 per count [11]; the server ranks triple candidates by the proximity of estimated support (products of sketched marginals) to the band centre. Round 2 queries the top K1 = 800 triple candidates together with all previously unsampled pairs, so pair coverage is complete. Round 3 expands the neighbourhood of near-band findings with K2 = 500 further candidates. Round 4 re-verifies borderline candidates (estimated support in [0.4%, 2.4%]): supporting clients self-select through the skip rule (noisy sketch count at least 1) and re-report at epsilon = 16 inside one secure-aggregated message [7]. Communication is the total pickled bytes of every uplink and downlink message (numpy uint32 masks and float32 counts); leakage is the per-client itemset and support content visible to the server, which for the plaintext baselines is their entire uplink and for the federated sketch protocols is only the verification participation pattern (one bit per borderline candidate per client). This realizes the Stage-A/B/C architecture of Section 4 with simplifications disclosed here and revisited in Section 5.6: candidate pooling uses a sampled noisy sketch rather than hashed local digests; the skip rule uses a fixed threshold tau = 1 rather than the skew-adaptive thresholds of *Eq. (5)*; the admission controller of *Eq. (4)* was not exercised (round budgets were fixed a priori); and the rule confidence gamma of *Eq. (2)* is not applied, band membership being by support alone.

The source reports Python 3.12.15, NumPy 2.1.3, SciPy 1.14.1, statsmodels 0.14.5, and Matplotlib 3.9.2, a 2-core Xeon with 3 GB RAM, and 61.5 s for the main campaign. The summaries are reported as means plus or minus standard deviations over twenty alpha-seed cells unless stated otherwise. Artifact-level confirmation of seeds, versions, runs, and formatting requires the experiment script and raw outputs.

5.2 | Main Comparison (RQ1)

Table 3 gives mean recall 0.9564 for FedRPM and 0.9234 for naive broadcast, a 3.30 percentage-point difference, with a reported paired Wilcoxon W=6 and p=2.67e-05. Serialized communication is 67.22 versus 72.49 KB, a 7.3% reduction relative to naive broadcast. The full-lattice reference reaches recall 0.9926 at 82.59 KB: 3.62 percentage points more recall and 22.9% more communication relative to FedRPM; equivalently, FedRPM uses 18.6% fewer bytes relative to that reference. The transmitted-content surrogate is 478.1 versus 74225.2 bytes, a ratio of about 155, but this is not a quantified privacy loss. Local-only has recall 0.9878 and precision 0.3530; these metrics imply about 2.80 times as many reported patterns as true-band patterns when interpreted at the aggregate-summary level.

Table 3. Reported mean and standard deviation over twenty simulated federations.

Method	Rare-Band Recall	Precision	Communication (KB)	Leakage (Bytes)
Centralized Apriori (oracle)	1.0000 ± 0.0000	1.0000 ± 0.0000	31.40 ± 0.00	32152.0 ± 0.0
Local-only (1% local min-sup)	0.9878 ± 0.0097	0.3530 ± 0.0084	60.08 ± 2.94	61522.0 ± 3006.0
Naive broadcast (0.5% local)	0.9234 ± 0.0136	0.9896 ± 0.0050	72.49 ± 2.63	74225.2 ± 2696.4
Full-lattice federated	0.9926 ± 0.0028	0.9999 ± 0.0004	82.59 ± 0.20	487.6 ± 35.4
FedRPM (proposed)	0.9564 ± 0.0144	0.9999 ± 0.0004	67.22 ± 0.23	478.1 ± 34.9

Fig. 1 compares reported recall and serialized communication. FedRPM has higher recall and lower communication than naive broadcast at both skew levels; the full-lattice reference has higher recall and higher communication. On the floor-adjacent band, reported recalls are 0.929 for FedRPM and 0.621 for naive broadcast. A globally supported pattern cannot fall below the same threshold at every client, since global support is a weighted average. These losses may instead reflect omitted low-support client contributions, incomplete candidate expansion, or the counting implementation; raw outputs are required to isolate the cause. FedRPM pair and triple recall are reported as 0.995 and 0.955.

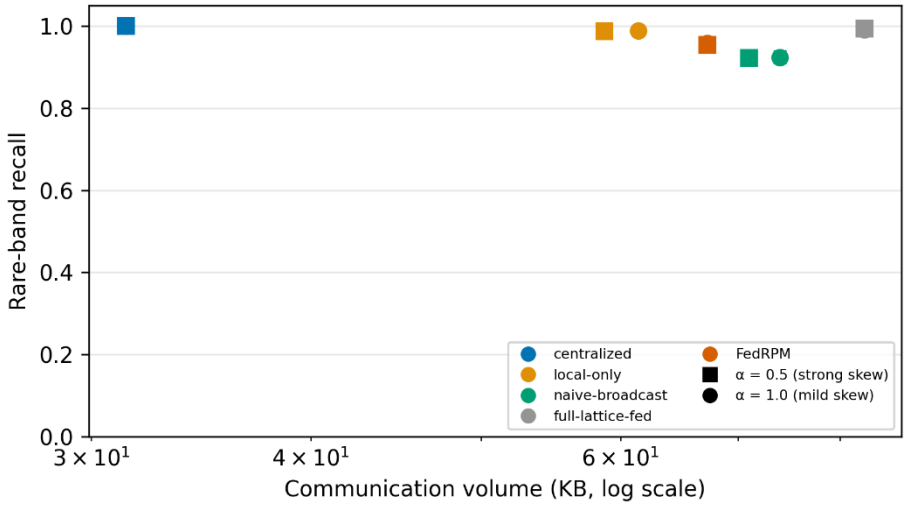


Fig. 1. Reported rare-band recall versus serialized communication, with one-standard-deviation error bars over ten seeds per skew level.

FedRPM has higher recall and lower bytes than naive broadcast; the full-lattice reference uses 22.9% more bytes than FedRPM for about 3.6 percentage points more recall. Secure aggregation is simulated.

Per-skew means show recall of 0.9549 versus 0.9227 at alpha = 0.5 and 0.9579 versus 0.9242 at alpha = 1.0 for FedRPM and naive broadcast. *Table 4* gives serialized communication of 67.22 and 67.23 KB for FedRPM, compared with 71.02 and 73.95 KB for naive broadcast. The exposure surrogate rises from 458.5 to 497.8 bytes. These conditional summaries do not isolate threshold adaptation or establish general robustness to skew.

Table 4. Reported per-skew-level means and standard deviations over ten seeds.

Alpha	Method	Recall	Precision	Comm (KB)	Leak (Bytes)
0.5	Centralized Apriori (oracle)	1.0000 ± 0.0000	1.0000 ± 0.0000	31.40 ± 0.00	32152.0 ± 0.0
0.5	Local-only (1% local min-sup)	0.9874 ± 0.0110	0.3517 ± 0.0092	58.75 ± 2.16	60158.4 ± 2206.9
0.5	Naive broadcast (0.5% local)	0.9227 ± 0.0123	0.9886 ± 0.0043	71.02 ± 1.86	72727.2 ± 1909.5
0.5	Full-lattice federated	0.9934 ± 0.0027	1.0000 ± 0.0000	82.60 ± 0.23	467.7 ± 19.9
0.5	FedRPM (proposed)	0.9549 ± 0.0147	0.9998 ± 0.0006	67.22 ± 0.25	458.5 ± 20.4
1.0	Centralized Apriori (oracle)	1.0000 ± 0.0000	1.0000 ± 0.0000	31.40 ± 0.00	32152.0 ± 0.0
1.0	Local-only (1% local min-sup)	0.9883 ± 0.0075	0.3543 ± 0.0068	61.41 ± 2.86	62885.6 ± 2929.1
1.0	Naive broadcast (0.5% local)	0.9242 ± 0.0141	0.9906 ± 0.0053	73.95 ± 2.33	75723.2 ± 2383.3
1.0	Full-lattice federated	0.9919 ± 0.0026	0.9998 ± 0.0006	82.58 ± 0.16	507.5 ± 34.6
1.0	FedRPM (proposed)	0.9579 ± 0.0131	1.0000 ± 0.0000	67.23 ± 0.19	497.8 ± 33.6

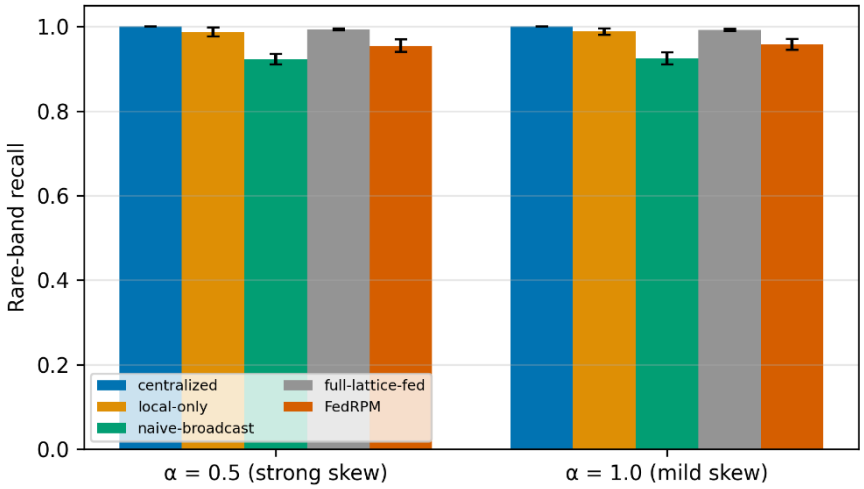


Fig. 2. Rare-band recall by Dirichlet skew level (mean plus one standard deviation over ten seeds).

FedRPM's advantage over naive broadcast holds at both alpha = 0.5 and alpha = 1.0; local-only attains high recall through permissive local thresholds but at the precision collapse documented in Table 3; the centralized oracle is perfect by construction.

5.3 | Ablation (RQ2, RQ3)

Table 5 reports an additional twenty-instance comparison for each variant. Removing verification lowers serialized communication from 67.24 to 59.32 KB and the participation-content surrogate from 478.7 to zero bytes; mean recall changes from 0.9574 to 0.9554 and precision from 0.9999 to 0.9993. No paired equivalence test is reported. Candidate budgets 400+250 and 1200+700 yield mean recalls 0.7470 and 0.9940 at 42.33 and 87.96 KB. Increasing sampled pairs from 80 to 240 changes recall from 0.9226 to 0.9771. These variants support sensitivity to candidate coverage; they do not isolate the effect of a correctly composed total privacy budget.

Table 5. Reported ablations over twenty alpha-seed cells per variant.

Variant	Recall	Precision	Comm (KB)	Leak (Bytes)	Candidates
FedRPM full (verify, K=800+500)	0.9574 ± 0.0104	0.9999 ± 0.0004	67.24	478.7	1393.0
No verification round	0.9554 ± 0.0120	0.9993 ± 0.0011	59.32	0.0	1393.0
Candidate budget K=400+250	0.7470 ± 0.0158	0.9998 ± 0.0007	42.33	361.5	743.0
Candidate budget K=1200+700	0.9940 ± 0.0029	0.9997 ± 0.0007	87.96	499.8	1970.9
Sketch: 80 sampled pairs	0.9226 ± 0.0120	0.9999 ± 0.0004	66.98	459.5	1473.0
Sketch: 240 sampled pairs	0.9771 ± 0.0057	0.9999 ± 0.0004	67.23	481.8	1313.0

K denotes round-2 and round-3 candidate budgets; sampled pairs come from a 253-pair universe.

5.4 | Per Coordinate Noise Sweep

Table 6 varies nominal per-count epsilon with verification epsilon_v=max(2 epsilon,1). These are coordinate noise parameters: Laplace(1/epsilon) protects a single exact count under add/remove adjacency, but repeated coordinates and rounds compose, and additional releases must be charged. Recall is 0.9221 at epsilon=0.5, 0.9485 at 1, 0.9579 at 4, and 0.9574 at 8. This smooth empirical noise response is not a total-epsilon privacy-utility frontier and does not validate *Eq. (4)*, whose E_count has a different meaning.

Table 6. Reported per-coordinate noise sweep over twenty simulated federations per setting.

Epsilon (Sketch)	Epsilon_V (Verify)	Recall	Precision
0.5	1	0.9221 ± 0.0144	0.9806 ± 0.0061
1	2	0.9485 ± 0.0140	0.9919 ± 0.0043
4	8	0.9579 ± 0.0107	0.9980 ± 0.0019
8	16	0.9574 ± 0.0104	0.9999 ± 0.0004

Epsilon and epsilon_v specify count-noise scales and do not represent an audited full-transcript privacy budget.

5.5 | Statistical Analysis

The source reports Friedman and two-sided paired Wilcoxon tests over twenty alpha-seed cells, with Holm correction for the four recall contrasts [28–30]. Friedman chi-square(4)=75.66 and p=1.4e-15 summarize variation across the five methods in this resampling design. *Table 7* retains the reported pairwise results; naive broadcast is the sole recall comparator favouring FedRPM, while the oracle, local-only, and full-lattice reference have higher recall. These cells are repeated partitions and subsamples of one corpus, not twenty independent datasets. Preregistration and the exact handling of ties, zeros, and shared seeds are not documented here; the tests are exploratory conditional summaries requiring raw per-run reconciliation.

Table 7. Pairwise two-sided Wilcoxon signed-rank tests of FedRPM against each baseline on per-run rare-band recall (twenty paired runs), with Holm correction over the four comparisons.

Comparison (Rare-Band Recall)	W	p	Holm p	Outcome
FedRPM vs. centralized	0.0	1.91e-06	7.63e-06	Centralized oracle better
FedRPM vs. local only	0.0	1.91e-06	7.63e-06	Local-only better (recall)
FedRPM vs. naive broadcast	6.0	2.67e-05	2.67e-05	FedRPM better
FedRPM vs. full lattice fed	0.0	1.91e-06	7.63e-06	Full-lattice reference better

5.6 | Findings, Honest Losses, and Scope

The campaign shows a favourable recall-communication point for FedRPM relative to naive broadcast at the measured settings, with simulated confidentiality and a smaller server-visible per-client content surrogate. It misses 4.36% of the reference band and has lower recall than the full-lattice reference and local-only union. The latter has much lower precision. FedRPM is also more expensive than local-only by the serialized-byte measure. Verification trades additional bytes for a small mean precision increase; its reported mean recall change does not establish equivalence or absence of cost.

H1 has empirical support only for the measured comparison with naive broadcast, not for a universal frontier dominance or asymptotic sublinear scaling. The 23-item universe is small and pair coverage completes in round 2. H2 remains open as a whole-budget privacy hypothesis: there is no secure-aggregation-only arm, and the sweep uses per-coordinate epsilon. H3 is untested because the campaign uses a fixed tau=1 rather than *Eq. (5)*.

The evidence is restricted to UCI Adult, 23 items, 8,000 baskets per run, eight clients, two skew levels, ten seeds per level, and support-band itemsets of size two or three. The average band contains 545.5 itemsets, including six planted carrier-concentrated patterns. Digests, fixed-threshold skipping, and aggregate interfaces are simulated; no cryptographic implementation, dropout study, adaptive-threshold test, or full-transcript

privacy proof is provided. Epsilon is per queried count, participation information is observable, and preprocessing and candidate discovery require accounting. The rule-confidence criterion is not evaluated.

The source reports a proof-of-concept campaign dated 2026-10-11. Its results remain conditional on the documented simulation and require experiment scripts, raw outputs, and generated figures for independent reproduction.

6 | Discussion: Limitations and Counterarguments

6.1 | Limitations

The Gaussian screen in *Eq. (4)* assumes independent Laplace noise, known normalization, and a dedicated count budget. The campaign uses different per-coordinate noise settings and does not exercise the controller, so it cannot validate that inequality. Skipping, selection, preprocessing, and transcript metadata prevent an end-to-end privacy claim without further accounting. The Dirichlet partitions model one form of heterogeneity; active attacks, collusion, dropouts, production overhead, and institutional-level privacy are outside the evidence.

6.2 | Alternative Viewpoints and Responses

Authorized pooling and contractual data sharing may be appropriate when legal and organisational conditions permit them. Federation changes where records reside, while differential privacy, when established for all releases, addresses output-based inference. Synthetic data can also be released under differential privacy; preserving rare structure poses a utility challenge rather than proving that every such release necessarily violates privacy.

Pattern discovery returns explicit co-occurrences useful for exploratory oversight and hypothesis generation. Their supports here are noisy estimates rather than certified bounds. Syntactic anonymity has different protection semantics. The measured comparison is a simulation of candidate coverage, bytes, and exposed per-client content; these quantities should not be described as the cost of an established formal privacy guarantee.

7 | Conclusion and Future Directions

FedRPM combines bounded candidate screening, noisy pooled counts, and focused verification. In the reported twenty-federation UCI Adult simulation it reaches recall 0.9564, precision 0.9999, and 67.22 KB of serialized communication, versus 0.9234, 0.9896, and 72.49 KB for naive broadcast. The full-lattice reference reaches recall 0.9926 at 82.59 KB, while candidate-budget ablations span recall 0.7470 to 0.9940. A nominal per-coordinate epsilon sweep reaches recall 0.9221 at epsilon=0.5. These results concern a sampled-count realization; they do not establish a complete record-level privacy guarantee, cryptographic confidentiality, or statistical certification of individual band decisions.

Further evaluation should reserve and audit a transcript-wide budget for preprocessing, discovery, counting, verification, and metadata, and release the experiment artifacts. It should then compare fixed and adaptive thresholds, test a secure-aggregation-only arm, assess support-margin uncertainty and missing-contribution bias, equalize baseline counting semantics, and scale the item universe and federation. Cryptographic implementation, dropout and adversarial tests, and an institutional-level privacy variant remain separate development tasks.

Author Contributions

J. G.: study design, methodology, and original drafting. Sh. D. G T.: implementation, experiments, analysis, and review, editing. All authors approved the final version.

Code Availability

Reproduction requires exp_22.py, exp_22_results.json, the seed list, and the original figure-generating workflow. A public repository or attached supplement is not identified in this manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

UCI Adult is publicly available [26]. The reported partitions and planted patterns require the authors' seeded generation procedure and raw per-run results, including exp_22_results.json; a repository or supplement containing those artifacts must be provided to reproduce the reported campaign.

Conflicts of Interest

The authors declare no conflicts of interest.

Generative AI Disclosure

AI-assisted drafting tools were used in preparing the manuscript text under the direction and scientific responsibility of the authors; all methodological choices, claims, experimental results, and final content were determined and verified by the authors.

References

[1] Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 acm sigmod international conference on management of data* (pp. 207–216). Association for Computing Machinery. https://doi.org/10.1145/170035.170072

[2] Szathmary, L., & Amedeo Napoli, P. V. (2007). Towards rare itemset mining. *19th IEEE international conference on tools with artificial intelligence(ictai 2007)* (pp. 305–312). HALL Open Science. https://doi.org/10.1109/ICTAI.2007.30

[3] Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of the20th vldbconference* (pp. 487–499). Association for Computing Machinery. https://dl.acm.org/doi/10.5555/645920.672836

[4] Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *Proceedings of the 2000 ACM sigmod international conference on management of data* (pp. 1–12). Association for Computing Machinery. https://doi.org/10.1145/342009.335372

[5] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. Y. (2017). Communication-efficient learning of deep networks from decentralized data. *The 20th international conference on artificial intelligence and statistics* (pp. 1273–1282). PMLR. https://proceedings.mlr.press/v54/mcmahan17a.html

[6] Konecny, J., McMahan, H. B., Yu, F. X., Richtarik, P., Suresh, A. T., & Bacon., D. (20216). *Federated learning: Strategies for improving communication efficiency*. https://arxiv.org/html/1610.05492v2

[7] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., … & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. *Proceedings of the 2017 ACM sigsac conference on computer and communications security* (pp. 1175–1191). Association for Computing Machinery. https://doi.org/10.1145/3133956.3133982

[8] Kairouz, P., & McMahan, H. B. (2021). Advances and open problems in federated learning. In *Foundations and trends® in machine learning* (pp. 1–210). Now Publishers Inc. https://doi.org/10.1561/2200000083

[9] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, *37*, 50–60. https://doi.org/10.1109/MSP.2020.2975749

[10] Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). *Federated learning with Non-IID Data*. https://arxiv.org/abs/1806.00582

[11] Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and trends®in theoretical computer science*, *9*(3–4), 211–487. https://doi.org/10.1561/0400000042

[12] Dong, G., & Li, J. (1999). Efficient mining of emerging patterns: discovering trends and differences. *Proceedings of the fifth acm sigkdd international conference on knowledge discovery and data mining* (pp. 43–52). Association for Computing Machinery. https://doi.org/10.1145/312129.312191

[13] Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. *Proceedings of machine learning and systems*, *2*, 429–450. https://proceedings.mlsys.org/paper/2020/hash/1f5fe83998a09396ebe6477d9475ba0c-Abstract.html

[14] Cormode, G., & Muthukrishnan, S. (2005). An improved data stream summary: The count-min sketch and its applications. *Journal of algorithms*, *55*(1), 58–75. https://doi.org/10.1016/j.jalgor.2003.12.001

[15] Shitharth, S., Manoharan, H., Shankar, A., Alsowail, R. A., Pandiaraj, S., Edalatpanah, S. A., & Viriyasitavat, W. (2023). Federated learning optimization: A computational blockchain process with offloading analysis to enhance security. *Egyptian informatics journal*, *24*(4), 100406. https://doi.org/10.1016/j.eij.2023.100406

[16] Agrawal, R., & Srikant, R. (2000). Privacy-preserving data mining. *Proceedings of the 2000 ACM sigmod international conference on management of data* (pp. 439–450). Association for Computing Machinery. https://doi.org/10.1145/342009.335438.

[17] Lindell, Y., & Pinkas, B. (2000). Privacy preserving data mining. *Annual international cryptology conference* (pp. 36–54). Springer. https://doi.org/10.1007/3-540-44598-6_3

[18] Vaidya, J., & Clifton, C. (2002). Privacy preserving association rule mining in vertically partitioned data. *Proceedings of the eighth acm sigkdd international conference on knowledge discovery and data mining* (pp. 639–644). Association for Computing Machinery. https://doi.org/10.1145/775047.775142

[19] Kantarcioglu, M., & Clifton, C. (2004). Privacy-preserving distributed mining of association rules on horizontally partitioned data. *IEEE transactions on knowledge and data engineering*, *16*(9), 1026–1037. https://doi.org/10.1109/TKDE.2004.45

[20] Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y., & Theodoridis, Y. (2004). State-of-the-art in privacy preserving data mining. *ACM sigmod record*, *33*(1), 50–57. https://doi.org/10.1145/974121.974131

[21] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. *Proceedings of the 2016 ACM sigsac conference on computer and communications security* (pp. 308–318). Association for Computing Machinery. https://doi.org/10.1145/2976749.2978318

[22] Erlingsson, Ú., Pihur, V., & Korolova, A. (2014). Rappor: randomized aggregatable privacy-preserving ordinal response. *Proceedings of the 2014 acm sigsac conference on computer and communications security* (pp. 1054–1067). Association for Computing Machinery. https://doi.org/10.1145/2660267.2660348

[23] Duchi, J. C., Jordan, M. I., & Wainwright, M. J. (2013). Local privacy and statistical minimax rates. *2013 ieee 54th annual symposium on foundations of computer science* (pp. 429–438). IEEE. https://doi.org/10.1109/FOCS.2013.53

[24] Wang, T., Li, N., & Jha, S. (2018). Locally differentially private frequent itemset mining. *2018 IEEE symposium on security and privacy (SP)* (pp. 127–143). IEEE. https://doi.org/10.1109/SP.2018.00035

[25] Kairouz, P., Oh, S., & Viswanath, P. (2015). The composition theorem for differential privacy [presentation]. *International conference on machine learning* (pp. 1376–1385). PMLR. https://proceedings.mlr.press/v37/kairouz15.html

[26] Kohavi, B., & Becker, R. (1996). *Adult [Dataset]*. https://doi.org/10.24432/C5XW20

[27] Hsu, T.-M. H., Qi, H., & Brown, M. (2019). *Measuring the effects of non-identical data distribution for federated visual classification*. https://arxiv.org/abs/1909.06335

[28] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of machine learning research*, *7*(1), 1–30. http://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf

[29] Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American statistical association*, *32*, 675–701. https://doi.org/10.1080/01621459.1937.10503522

[30] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin*, *1*, 80. https://doi.org/10.2307/3001968