2026 Network and Distributed System Security Symposium (NDSS)

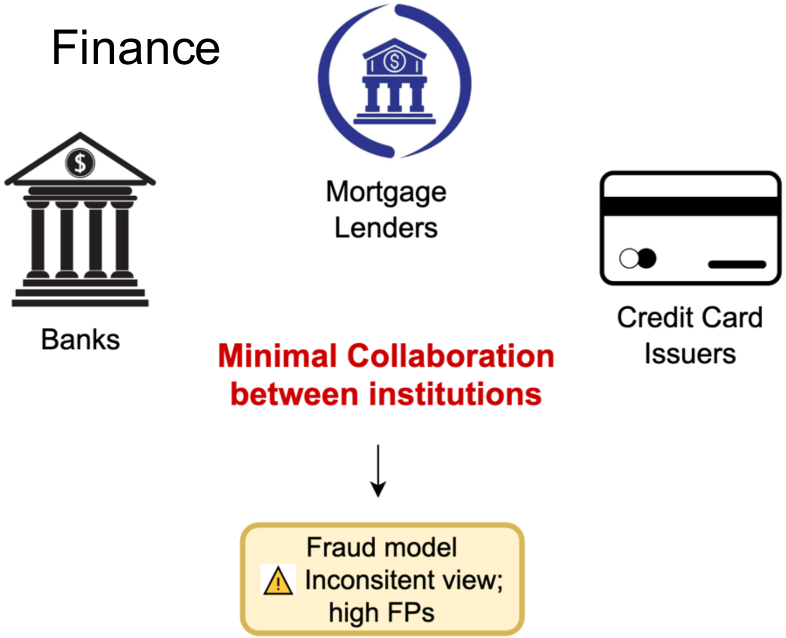
Select-Then-Compute: Encrypted Label Selection & Analytics over Distributed Datasets using FHE

Nirajan Koirala[1], Seunghun Paik[2], Sam Martin[1], Helena Berens[1], Tasha Januszewicz[1], Jonathan Takeshita[3], Jae Hong Seo[2], Taeho Jung[1]

[1]University of Notre Dame, [2]Hanyang University, [3]Old Dominion University

Current Data Landscape

Finance



Banks

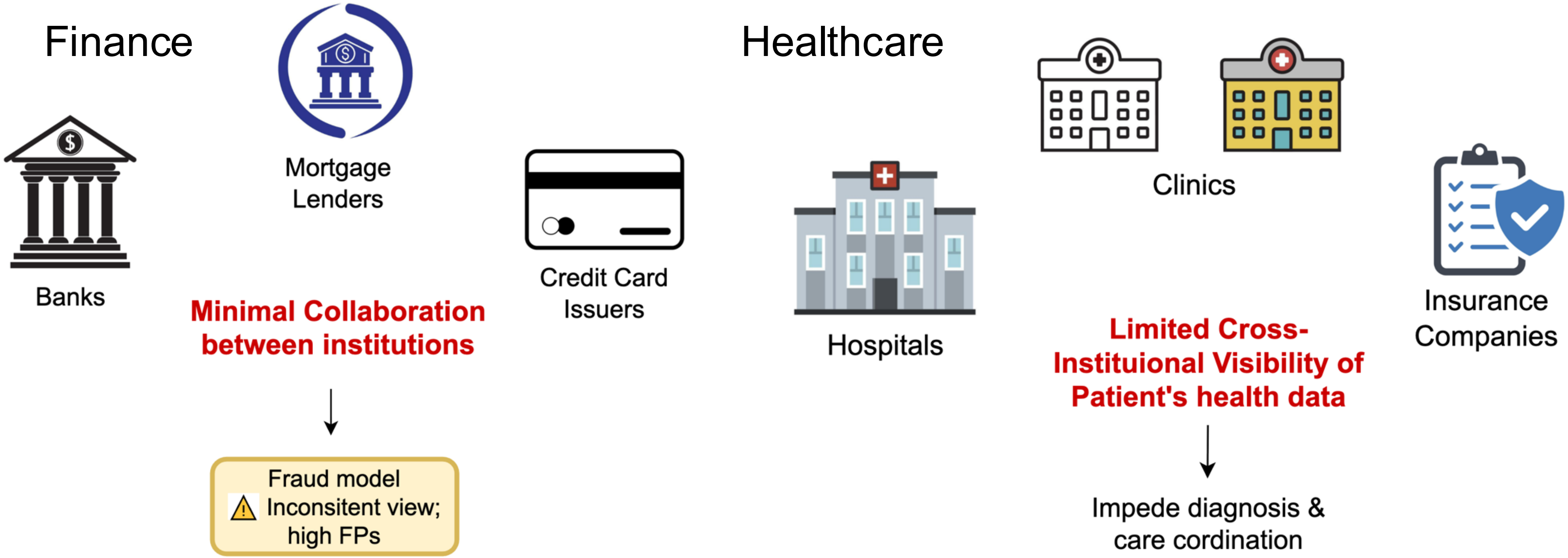
Mortgage Lenders

Credit Card Issuers

Minimal Collaboration between institutions

Fraud model
⚠ Inconsitent view; high FPs

Current Data Landscape



Finance

Banks

Mortgage Lenders

Credit Card Issuers

Minimal Collaboration between institutions

↓

⚠ Fraud model Inconsitent view; high FPs

Healthcare

Hospitals

Clinics

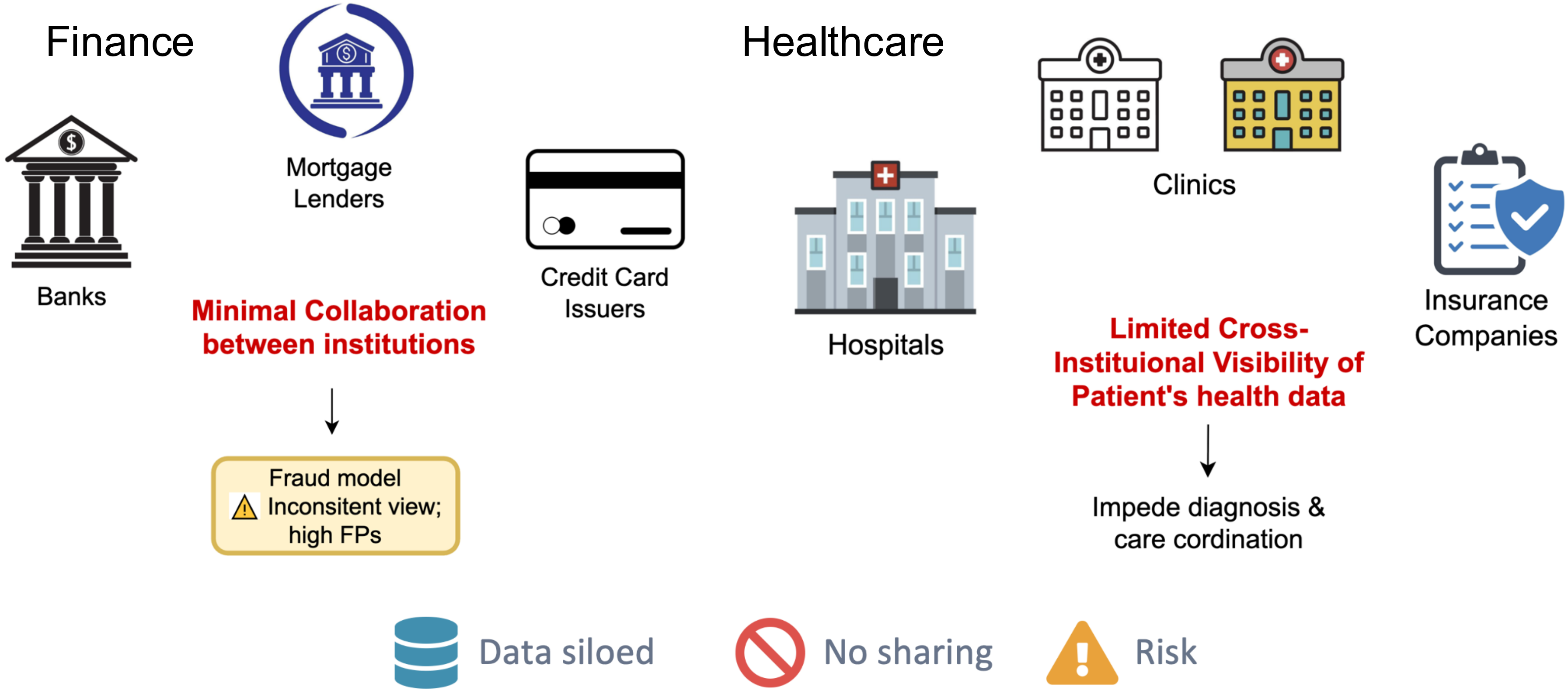
Insurance Companies

Limited Cross-Instituional Visibility of Patient's health data

↓

Impede diagnosis & care cordination

Current Data Landscape



Finance

Banks

Mortgage Lenders

Credit Card Issuers

Minimal Collaboration between institutions

↓

Fraud model ⚠ Inconsitent view; high FPs

Healthcare

Hospitals

Clinics

Insurance Companies

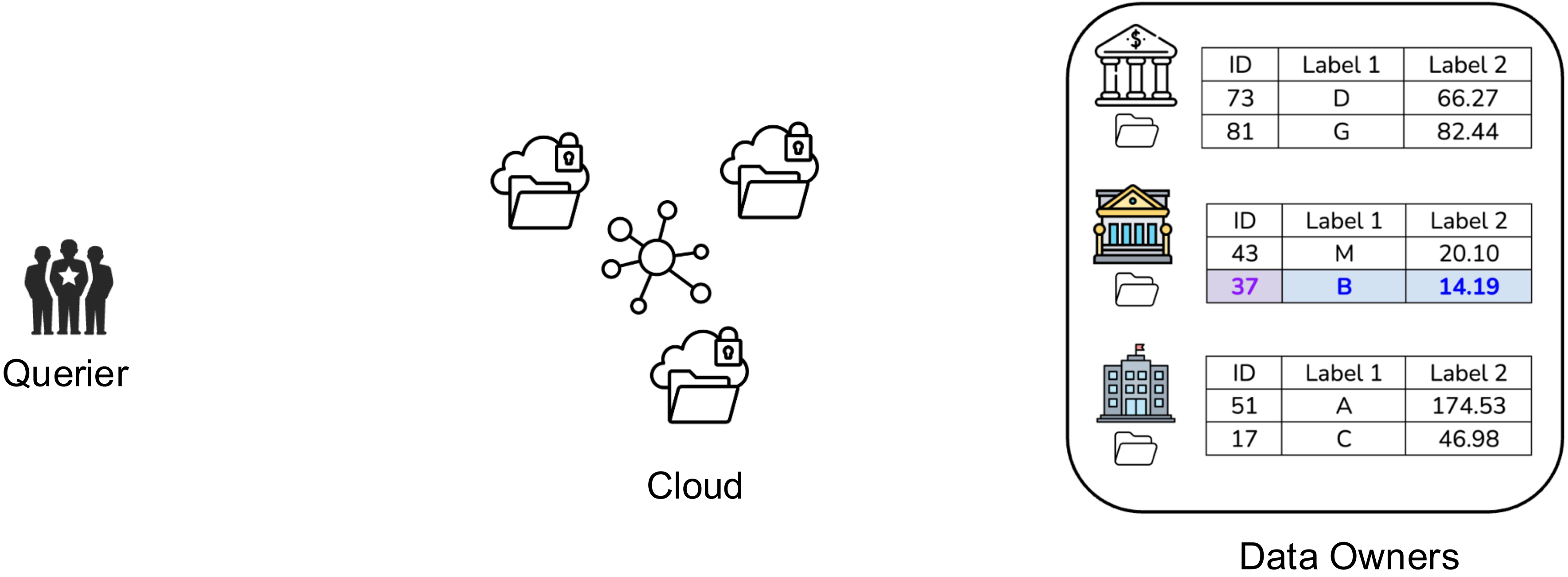
Limited Cross-Instituional Visibility of Patient's health data

↓

Impede diagnosis & care cordination

🛢 Data siloed ⊘ No sharing ⚠ Risk

Problem: Label Selection & Analytics

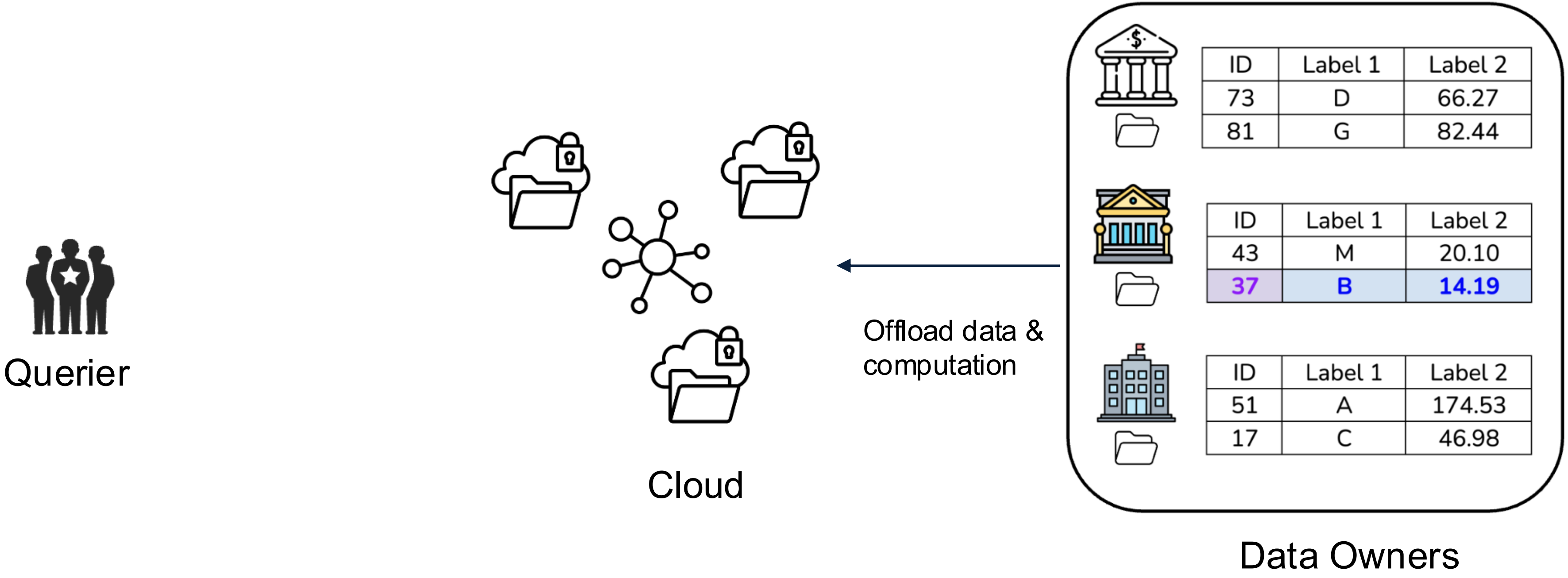


ID	Label 1	Label 2
73	D	66.27
81	G	82.44

ID	Label 1	Label 2
43	M	20.10
37	B	14.19

ID	Label 1	Label 2
51	A	174.53
17	C	46.98

Problem: Label Selection & Analytics



Querier

Cloud

Offload data & computation

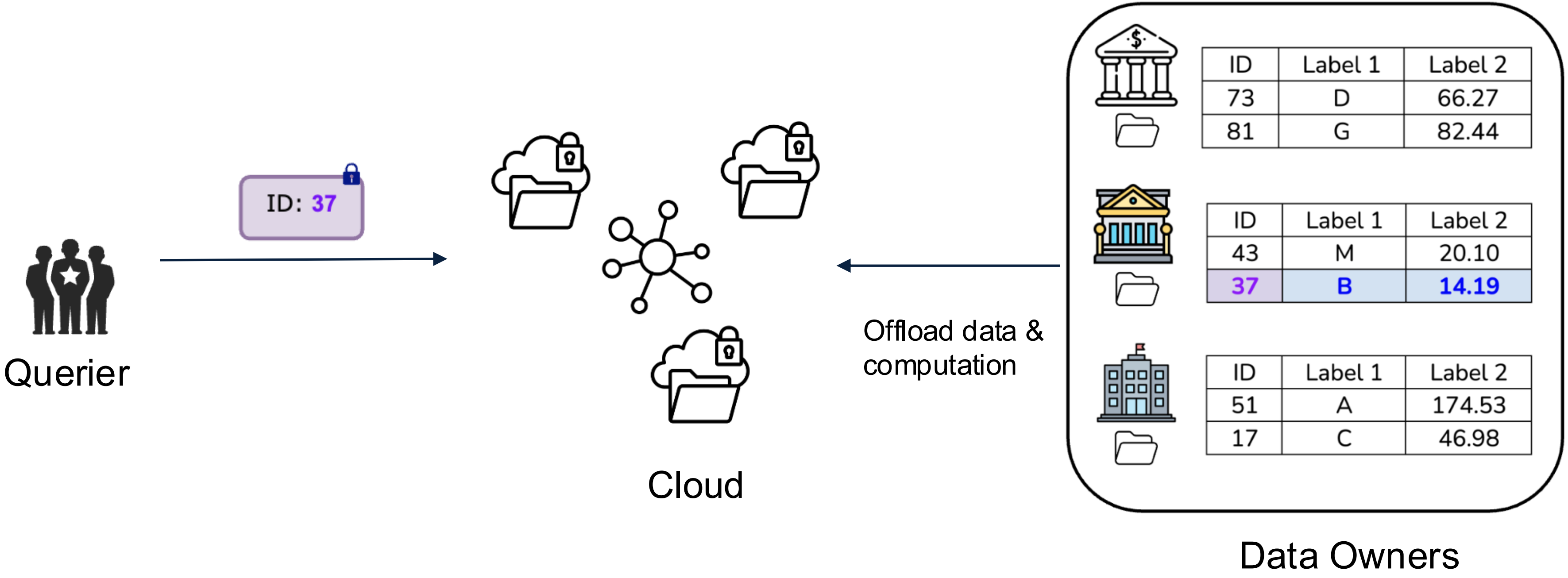
ID	Label 1	Label 2
73	D	66.27
81	G	82.44

ID	Label 1	Label 2
43	M	20.10
37	B	14.19

ID	Label 1	Label 2
51	A	174.53
17	C	46.98

Data Owners

Problem: Label Selection & Analytics

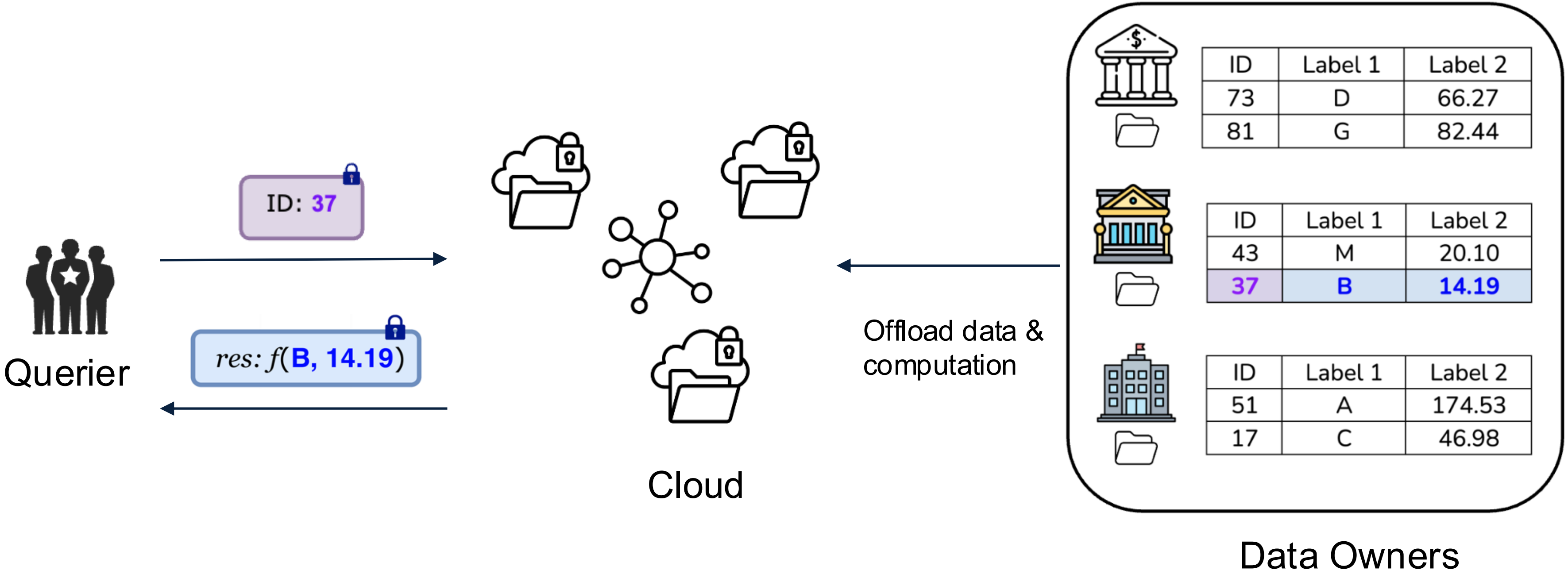


ID	Label 1	Label 2
73	D	66.27
81	G	82.44

ID	Label 1	Label 2
43	M	20.10
37	B	14.19

ID	Label 1	Label 2
51	A	174.53
17	C	46.98

Problem: Label Selection & Analytics

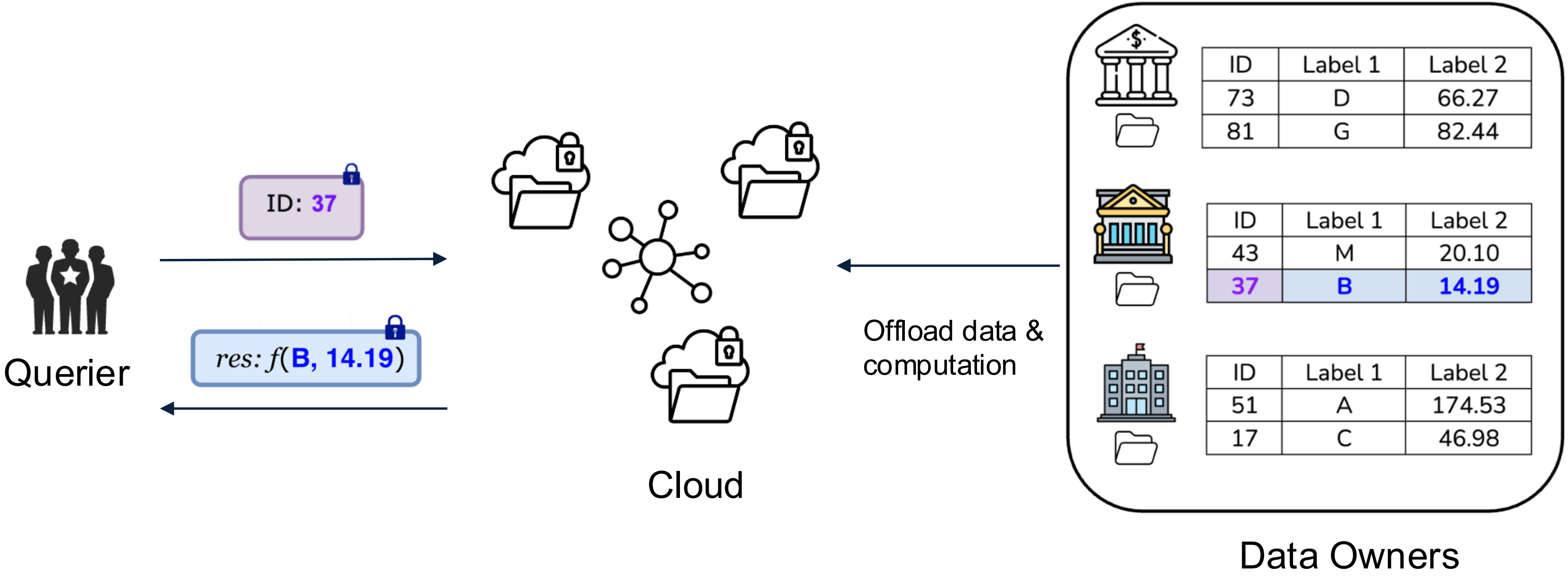


ID	Label 1	Label 2
73	D	66.27
81	G	82.44

ID	Label 1	Label 2
43	M	20.10
37	B	14.19

ID	Label 1	Label 2
51	A	174.53
17	C	46.98

Problem: Label Selection & Analytics



ID	Label 1	Label 2
73	D	66.27
81	G	82.44

ID	Label 1	Label 2
43	M	20.10
37	B	14.19

ID	Label 1	Label 2
51	A	174.53
17	C	46.98

Need: Privacy-preserving cross-institutional analytics on (encrypted) large ID-label pairs at scale

Relevant to Many Cryptographic Primitives

The problem intersects several well-studied cryptographic primitives , but none fully solves it.

Relevant to Many Cryptographic Primitives

The problem intersects several well-studied cryptographic primitives , but none fully solves it.

Private Set Intersection[1]

- Product-based PSI computes polynomial products over identifiers

Not inherently suitable:

- Designed for intersection, not associated label retrieval & analytics
- Extending to real-valued functions requires fundamental redesign

Private Information Retrieval[2]

- Retrieves records by index without revealing which index was queried

Limitations:

- Querier must know the exact index — no identifier-based lookup across holders
- No native support for post-retrieval computation on encrypted results

Private Segmented Membership Test[3]

- Similar to PSI but for singleton input sets and large-scale distributed datasets

Solves some but not all:

- Cannot approximate 1 and 0 exactly — only achieves value separation
- No label retrieval, or downstream analytics over real-values

[1] CCS 2021: Cong et al., Labeled PSI from homomorphic encryption with reduced computation and communication

[2] USENIX 2023: Henzinger et al., One Server for the Price of Two: Simple and Fast Single-Server Private Information Retrieval

[3] PETS 2024: Koirala et al., Summation-based private segmented membership test from threshold-fully homomorphic encryption

Relevant to Many Cryptographic Primitives

The problem intersects several well-studied cryptographic primitives , but none fully solves it.

Private Set Intersection[1]

- Product-based PSI computes polynomial products over identifiers

Not inherently suitable:

- Designed for intersection, not associated label retrieval & analytics
- Extending to real-valued functions requires fundamental redesign

Private Information Retrieval[2]

- Retrieves records by index without revealing which index was queried

Limitations:

- Querier must know the exact index — no identifier-based lookup across holders
- No native support for post-retrieval computation on encrypted results

Private Segmented Membership Test[3]

- Similar to PSI but for singleton input sets and large-scale distributed datasets

Solves some but not all:

- Cannot approximate 1 and 0 exactly — only achieves value separation
- No label retrieval, or downstream analytics over real-values

Each primitive addresses a piece of the problem, but it requires a unified, real-valued native solution

[1] CCS 2021: Cong et al., Labeled PSI from homomorphic encryption with reduced computation and communication

[2] USENIX 2023: Henzinger et al., One Server for the Price of Two: Simple and Fast Single-Server Private Information Retrieval

[3] PETS 2024: Koirala et al., Summation-based private segmented membership test from threshold-fully homomorphic encryption

Why existing methods fall short?

Protocol	Label Retrieval	Label Privacy	Real-Valued Function	Multi-Sender Scale	FHE-based Analytics
Labeled PSI [CHLR18, Cong+21]	✓	✗	✗	2-party	✗
Circuit-PSI [Son+23, Rindal+21]	✓	✓	✗	2-party	✗
PEPSI [Mahdavi+24]	✓	✓	✗	2-party	✓
MPSI [Wu+24, Nevo+21]	✗	—	—	Up to 100s	✗
KTSJ24 [Koirala+24]	✗	—	—	1000s+	✗
This Work	✓	✓	✓	**1000s+**	✓

CCS 2018: CHLR18, Chen et al., Labeled PSI from Fully Homomorphic Encryption with Malicious Security
CCS 2021: Cong et al., Labeled PSI from homomorphic encryption with reduced computation and communication
AsiaCCS 2023: Son et al., PSI with computation or Circuit-PSI for Unbalanced Sets from Homomorphic Encryption
EUROCRYPT 2021: Rindal et al., VOLE-PSI: fast OPRF and circuit-PSI from vector-OLE
USENIX 2024: Mahdavi et al., {PEPSI}: Practically Efficient Private Set Intersection in the Unbalanced Setting
USENIX 2024: Wu et al., {O-Ring} and {K-Star}: Efficient Multi-party Private Set Intersection
CCS 2021: Nevo et al., Simple, fast malicious multiparty private set intersection
PETS 2024: KTSJ24, Koirala et al., Summation-based private segmented membership test from threshold-fully homomorphic encryption

Why existing methods fall short?

Protocol	Label Retrieval	Label Privacy	Real-Valued Function	Multi-Sender Scale	FHE-based Analytics
Labeled PSI [CHLR18, Cong+21]	✓	✗	✗	2-party	✗
Circuit-PSI [Son+23, Rindal+21]	✓	✓	✗	2-party	✗
PEPSI [Mahdavi+24]	✓	✓	✗	2-party	✓
MPSI [Wu+24, Nevo+21]	✗	—	—	Up to 100s	✗
KTSJ24 [Koirala+24]	✗	—	—	1000s+	✗
This Work	✓	✓	✓	**1000s+**	✓

None supports: (1) encrypted label retrieval w/ privacy , (2) real-valued downstream analytics, (3) multi-sender scalability with minimal communication (FHE-based construction)

CCS 2018: CHLR18, Chen et al., Labeled PSI from Fully Homomorphic Encryption with Malicious Security
CCS 2021: Cong et al., Labeled PSI from homomorphic encryption with reduced computation and communication
AsiaCCS 2023: Son et al., PSI with computation or Circuit-PSI for Unbalanced Sets from Homomorphic Encryption
EUROCRYPT 2021: Rindal et al., VOLE-PSI: fast OPRF and circuit-PSI from vector-OLE
USENIX 2024: Mahdavi et al., {PEPSI}: Practically Efficient Private Set Intersection in the Unbalanced Setting
USENIX 2024: Wu et al., {O-Ring} and {K-Star}: Efficient Multi-party Private Set Intersection
CCS 2021: Nevo et al., Simple, fast malicious multiparty private set intersection
PETS 2024: Koirala et al., Summation-based private segmented membership test from threshold-fully homomorphic encryption

Breaking down the Problem

Breaking down the Problem

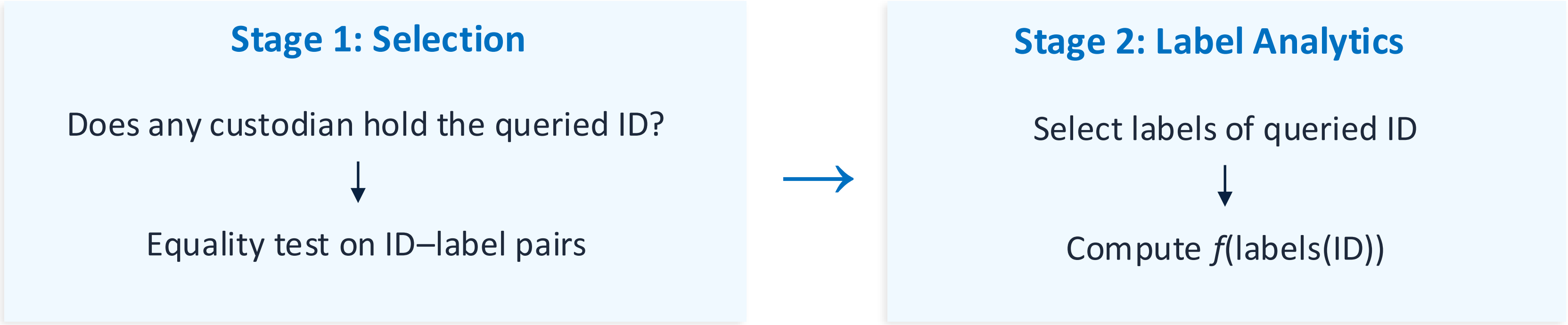
Stage 1: Selection

Does any custodian hold the queried ID?

↓

Equality test on ID–label pairs

Breaking down the Problem



Stage 1: Selection

Does any custodian hold the queried ID?

↓

Equality test on ID–label pairs

→

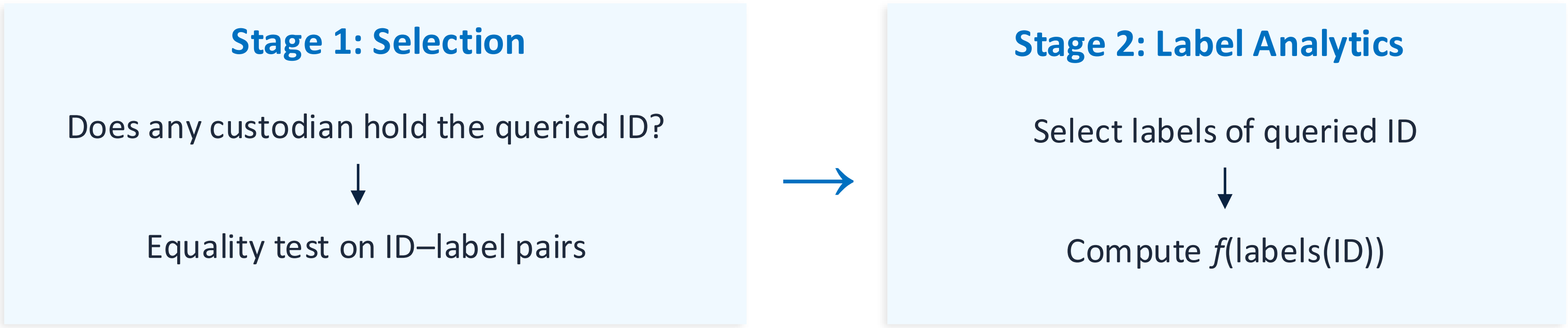
Stage 2: Label Analytics

Select labels of queried ID

↓

Compute *f*(labels(ID))

Breaking down the Problem



Stage 1: Selection

Does any custodian hold the queried ID?

↓

Equality test on ID–label pairs

→

Stage 2: Label Analytics

Select labels of queried ID

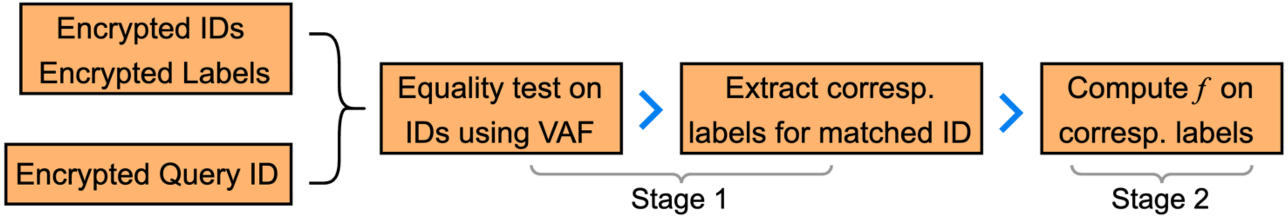
↓

Compute f(labels(ID))

Requirements

R1	**Label Confidentiality**	Labels remain encrypted end-to-end
R2	**Secure Analytics**	Complex computations on encrypted labels
R3	**Fragmentated Data**	Scale to large number of parties without centralizing data

Our Protocol at 10,000 ft



Encrypted IDs
Encrypted Labels

Encrypted Query ID

Equality test on IDs using VAF

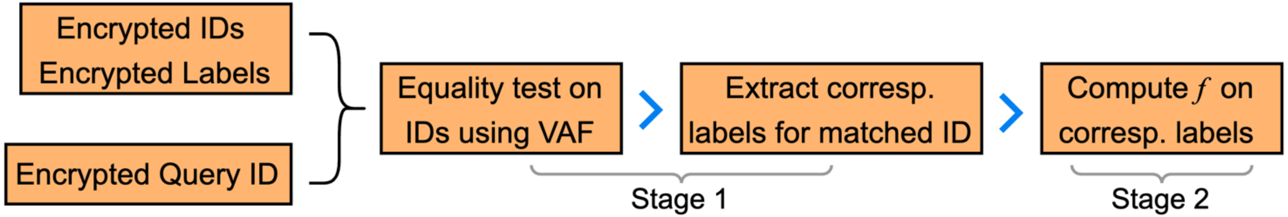
Extract corresp. labels for matched ID

Compute f on corresp. labels

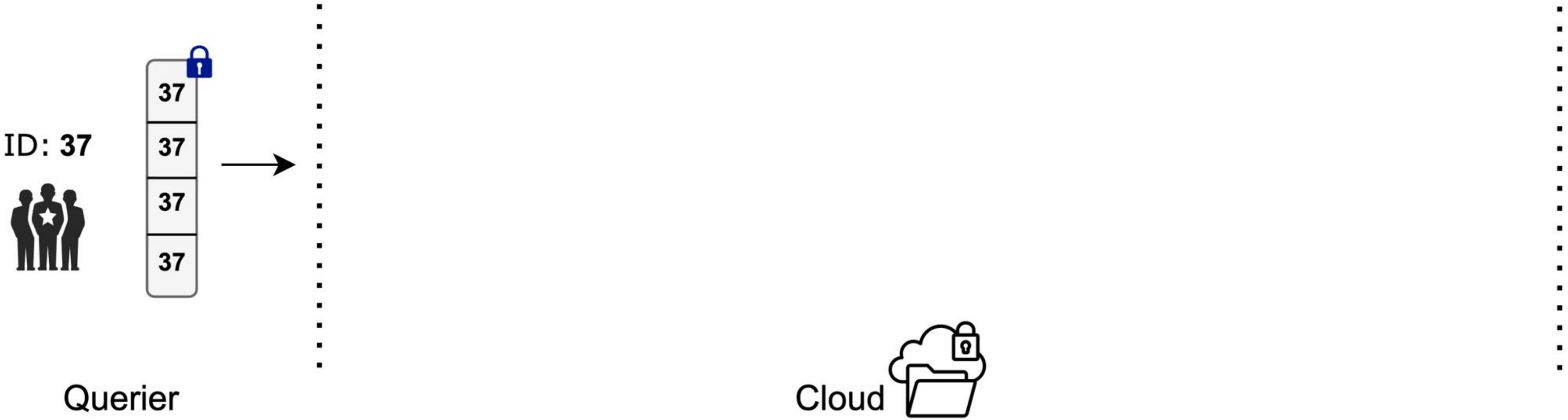
Stage 1

Stage 2

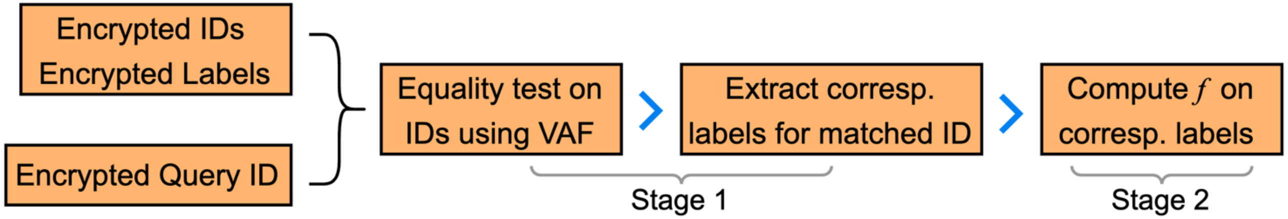
Our Protocol at 10,000 ft


Encrypted IDs
Encrypted Labels
Encrypted Query ID
Equality test on IDs using VAF
Extract corresp. labels for matched ID
Compute f on corresp. labels
Stage 1
Stage 2

Basic Idea



Our Protocol at 10,000 ft



Encrypted IDs
Encrypted Labels

Encrypted Query ID

Equality test on IDs using VAF

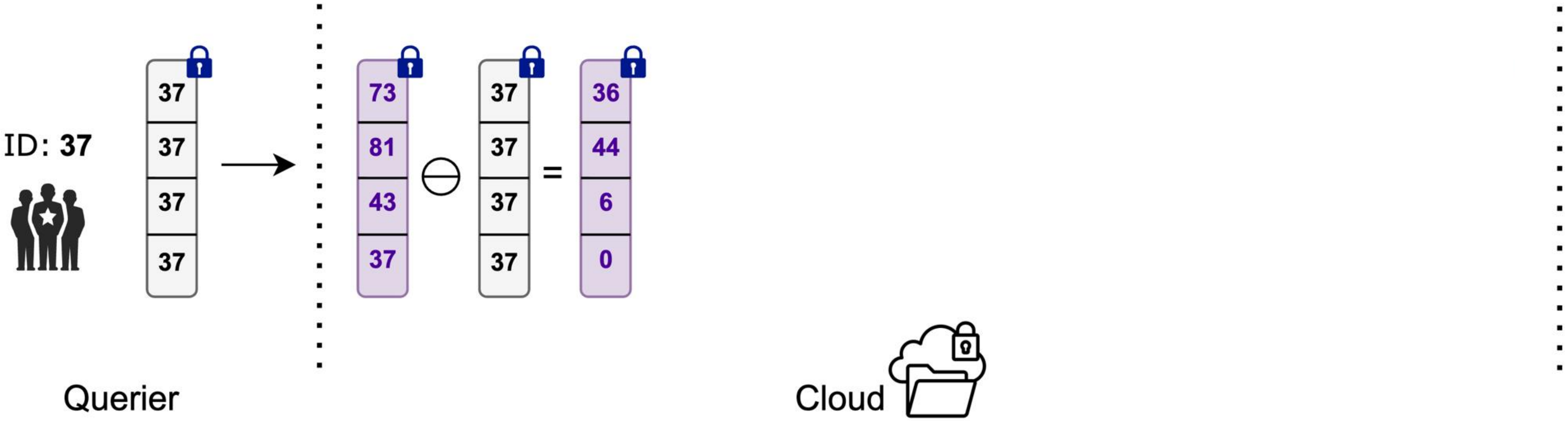
Extract corresp. labels for matched ID

Compute f on corresp. labels

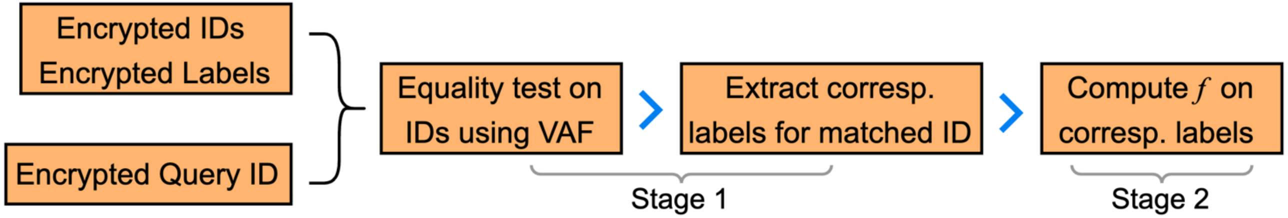
Stage 1

Stage 2

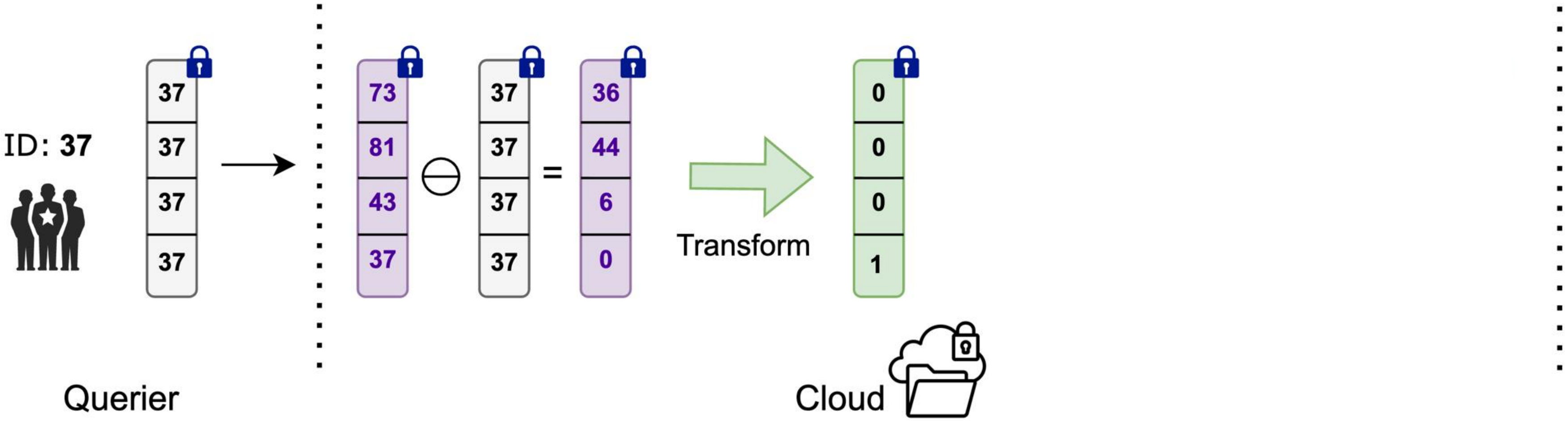
Basic Idea



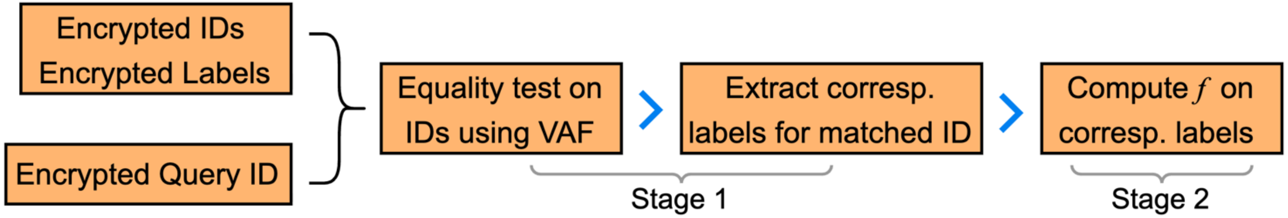
Our Protocol at 10,000 ft


Encrypted IDs
Encrypted Labels
Encrypted Query ID
Equality test on IDs using VAF
Extract corresp. labels for matched ID
Compute f on corresp. labels
Stage 1
Stage 2

Basic Idea



Our Protocol at 10,000 ft



Encrypted IDs
Encrypted Labels

Encrypted Query ID

Equality test on IDs using VAF

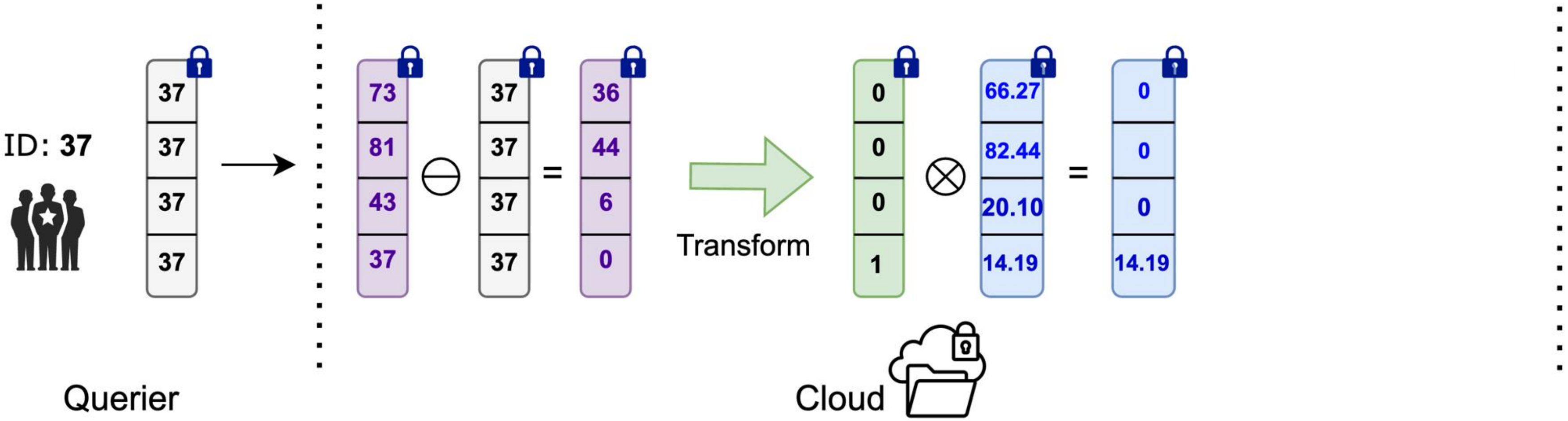
Extract corresp. labels for matched ID

Compute f on corresp. labels

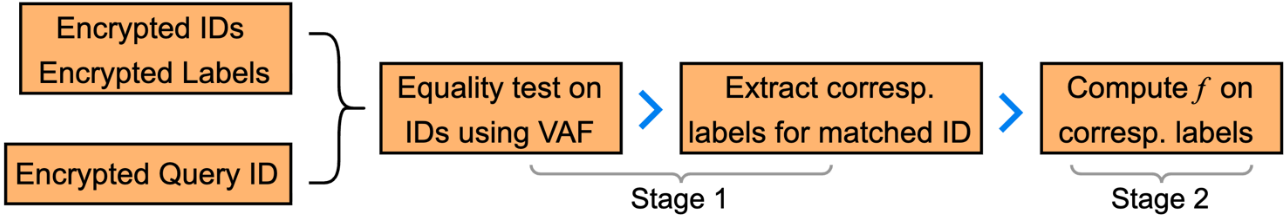
Stage 1

Stage 2

Basic Idea



Our Protocol at 10,000 ft



Encrypted IDs
Encrypted Labels

Encrypted Query ID

Equality test on IDs using VAF

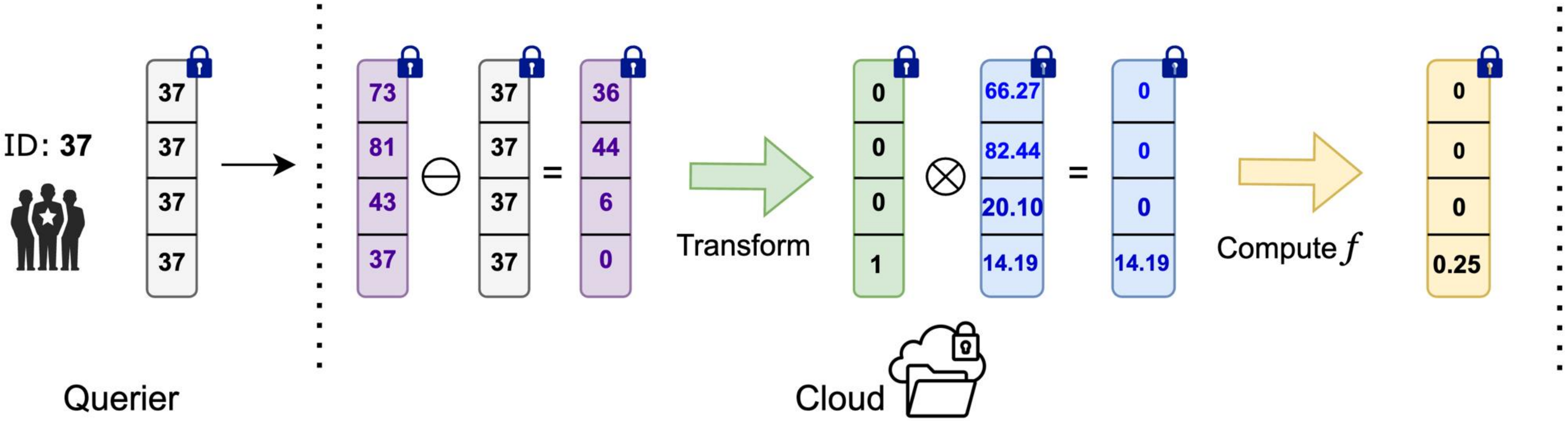
Extract corresp. labels for matched ID

Compute f on corresp. labels

Stage 1

Stage 2

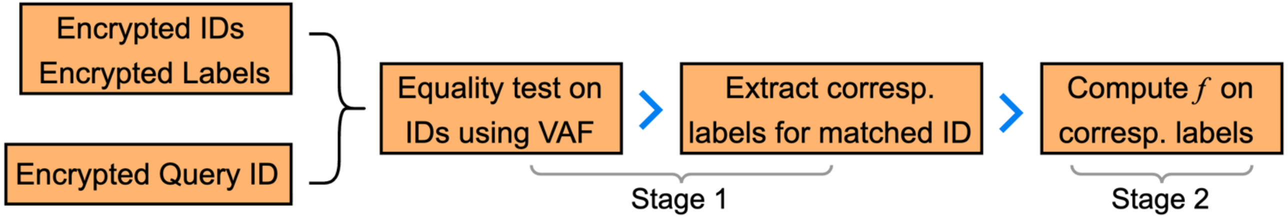
Basic Idea



Querier

Cloud

Our Protocol at 1,000 ft


Encrypted IDs
Encrypted Labels
Encrypted Query ID
Equality test on IDs using VAF
Extract corresp. labels for matched ID
Compute *f* on corresp. labels
Stage 1
Stage 2

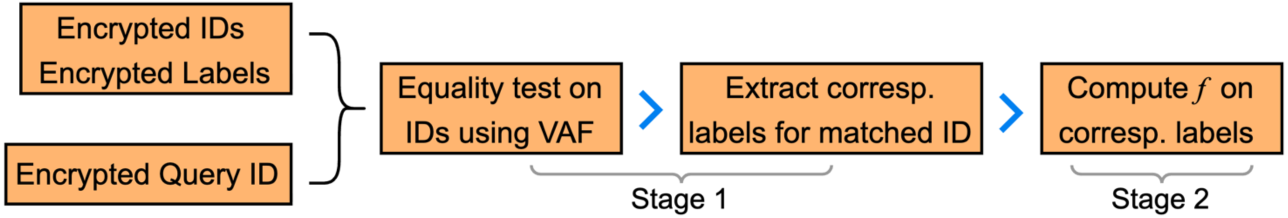
Stage 1 — Select & Extract

- Compute difference *d* between query ID & database IDs
- Evaluate homomorphic equality via VAF(d) ≈ $1\{d = 0\}$
- Multiply VAF * label → extract label for matching ID slot

Stage 2 — Compute *f* on labels

- Aggregate extracted labels into one ciphertext
- Evaluate analytic *f* (e.g., logistic regression / ML inference) and match flag
- Return result ciphertext + match flag to querier

Our Protocol at 1,000 ft


Encrypted IDs
Encrypted Labels
Encrypted Query ID
Equality test on IDs using VAF
Extract corresp. labels for matched ID
Compute f on corresp. labels
Stage 1
Stage 2

Stage 1 — Select & Extract

- Compute difference d between query ID & database IDs
- Evaluate homomorphic equality via VAF(d) ≈ $1\{d=0\}$
- Multiply VAF * label → extract label for matching ID slot

Stage 2 — Compute f on labels

- Aggregate extracted labels into one ciphertext
- Evaluate analytic f (e.g., logistic regression / ML inference) and match flag
- Return result ciphertext + match flag to querier

Novel VAF using wDEP ∘ Bell-shaped

- VAF: Value Annihilating Functions (f(x) = K if x = 0, else 0)
- Prior work[1] used DEPs (Domain Extension Polynomials) to compute VAFs
- **Limitations**: (1) Coarse grained approximation, (2) Do not offer label extraction

[1] PETS 2024: Koirala et al., Summation-based private segmented membership test from threshold-fully homomorphic encryption

Novel VAF using wDEP ∘ Bell-shaped

- VAF: Value Annihilating Functions (f(x) = K if x = 0, else 0)
- Prior work[1] used DEPs (Domain Extension Polynomials) to compute VAFs
- **Limitations**: (1) Coarse grained approximation, (2) Do not offer label extraction

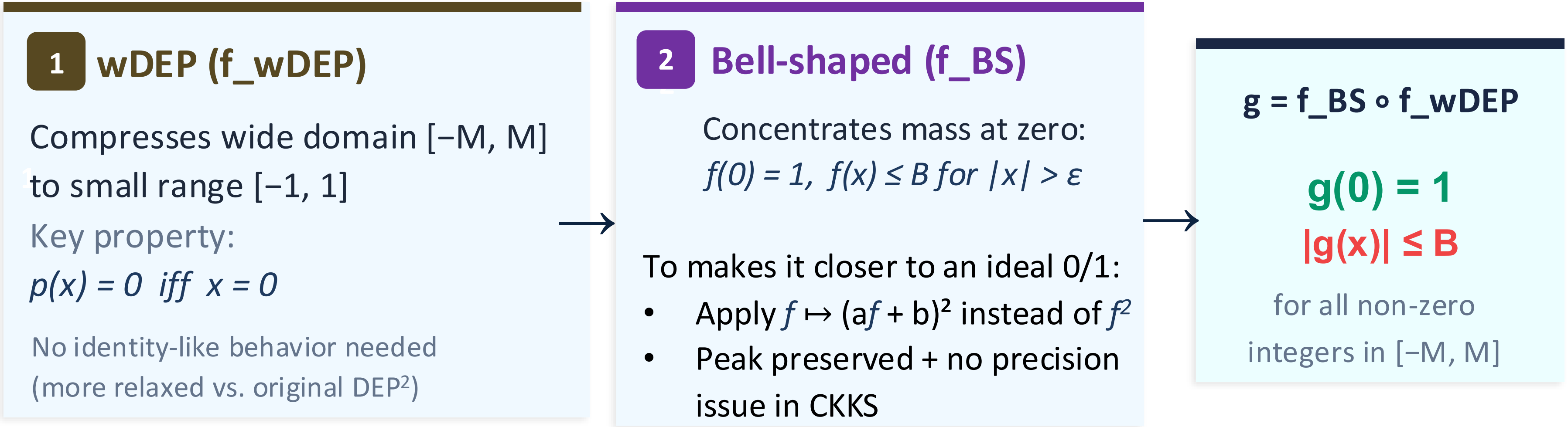
Our Idea: Approximate indicator f(x) = 1 if x = 0 with high fidelity under CKKS
→ *Compose* ***wDEP*** *+* ***Bell-shaped function***

[1] PETS 2024: Koirala et al., Summation-based private segmented membership test from threshold-fully homomorphic encryption

Novel VAF using wDEP ∘ Bell-shaped

- VAF: Value Annihilating Functions ($f(x) = K$ if $x = 0$, else 0)
- Prior work[1] used DEPs (Domain Extension Polynomials) to compute VAFs
- **Limitations**: (1) Coarse grained approximation, (2) Do not offer label extraction

Our Idea: Approximate indicator $f(x) = 1$ if $x = 0$ with high fidelity under CKKS
→ *Compose* ***wDEP*** + ***Bell-shaped function***



1 wDEP (f_wDEP)

Compresses wide domain [−M, M] to small range [−1, 1]

Key property:

$p(x) = 0 \ \text{iff} \ x = 0$

No identity-like behavior needed (more relaxed vs. original DEP[2])

→

2 Bell-shaped (f_BS)

Concentrates mass at zero:

$f(0) = 1, \ f(x) \leq B \text{ for } |x| > \varepsilon$

To makes it closer to an ideal 0/1:

- Apply $f \mapsto (af + b)^2$ instead of f^2
- Peak preserved + no precision issue in CKKS

→

g = f_BS ∘ f_wDEP

g(0) = 1

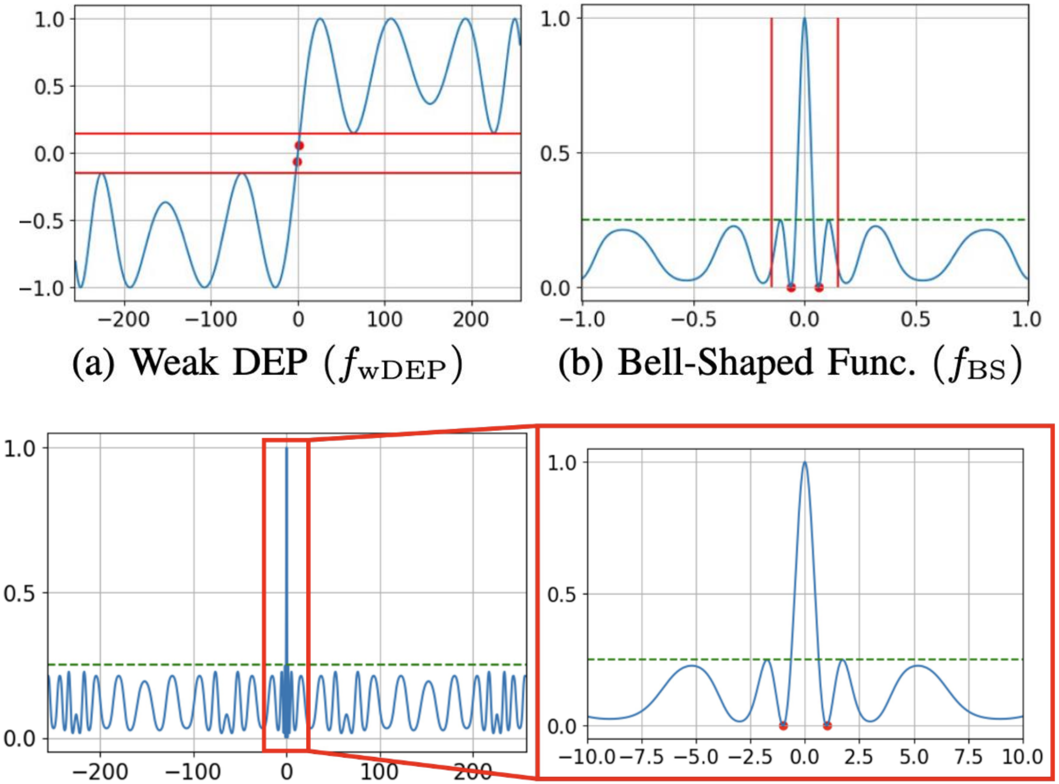
|g(x)| ≤ B

for all non-zero integers in [−M, M]

[1] PETS 2024: Koirala et al., Summation-based private segmented membership test from threshold-fully homomorphic encryption

[2] IEEE TIFS 2022: Cheon et al., Efficient homomorphic evaluation on large intervals

Novel VAF using wDEP ∘ Bell-shaped



(a) Weak DEP (f_{wDEP})

(b) Bell-Shaped Func. (f_{BS})

(c) Final VAF from Composition $(g = f_{\mathrm{BS}} \circ f_{\mathrm{wDEP}})$

Slot-Wise Windowing for large IDs

Challenge: Directly approximating VAF over 2^{64} or 2^{128} domains is prohibitively expensive in FHE

Slot-Wise Windowing for large IDs

Challenge: Directly approximating VAF over 2^{64} or 2^{128} domains is prohibitively expensive in FHE

Key Idea

Parse each δ-bit identifier into κ smaller windows of $[\delta/\kappa]$ bits each, test equality per chunk, then multiply:

$$x = 0 \iff x_1 = x_2 = \ldots = x_\kappa = 0 \iff \prod_{i=1}^{\kappa} f_{VAF,\xi}(x_i) = 1$$

κ parallel VAF evaluations + $\log_2(\kappa)$ depth for final multiplication

Slot-Wise Windowing for large IDs

Challenge: Directly approximating VAF over 2^{64} or 2^{128} domains is prohibitively expensive in FHE

Key Idea

Parse each δ-bit identifier into κ smaller windows of $[\delta/\kappa]$ bits each, test equality per chunk, then multiply:

$$x = 0 \Leftrightarrow x_1 = x_2 = \ldots = x_\kappa = 0 \Leftrightarrow \prod_{i=1}^{\kappa} f_{VAF,\xi}(x_i) = 1$$

κ parallel VAF evaluations + $\log_2(\kappa)$ depth for final multiplication

Method	κ	Storage Ciphertexts	Comm. (MB)	FHE Depth	Precision (bits)	Time (s)
KTSJ24	–	1	113.3	52	30.0	151.6
Ours (κ=1)	1	1	86.0	39	25.1	28.87
Ours (κ=5)	**5**	**5**	**37.0**	**16**	**39.6**	**5.60**
Ours (κ=10)	**10**	**10**	**33.0**	**14**	**42.1**	**5.48**

VAF over 2^{16} items, δ = 20 bits, single cloud server. κ = 5 or 10 achieves 27× speedup over KTSJ24 with higher precision

Evaluation & End-to-End latency

Setup

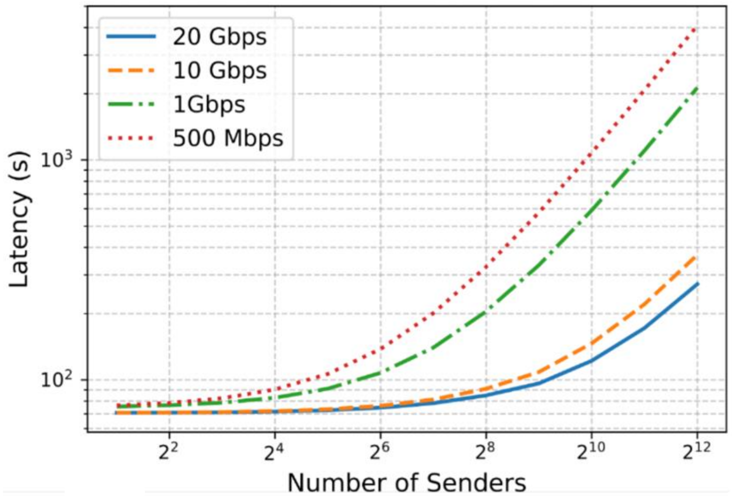
- Implementation: C++17 + OpenFHE v1.2.3 (threshold CKKS)
- Machine: Intel Xeon Gold 5412U, 512 GB RAM
- Default parameters for 128-bit classical security for FHE
- Downstream analytic: Logistic Regression Model

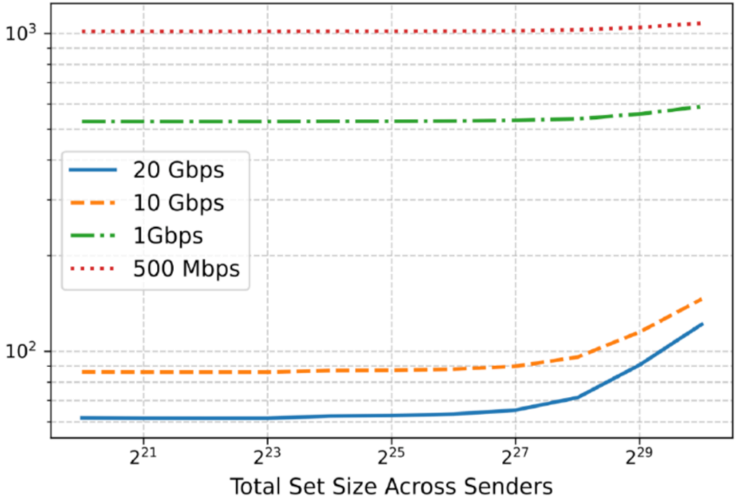
Evaluation

Two variants of our protocol:

1) Selection stage benchmarking and comparison with SOTA
2) Selection + Downstream Analytics on three fraud-oriented datasets (Fraud Dataset Benchmark)

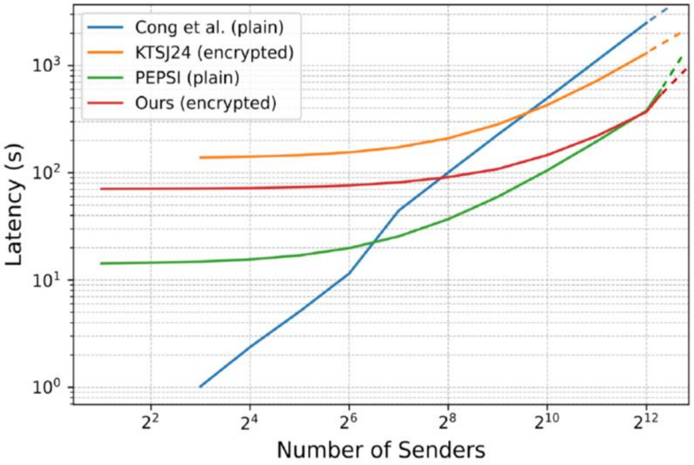
Selection Stage Benchmarking

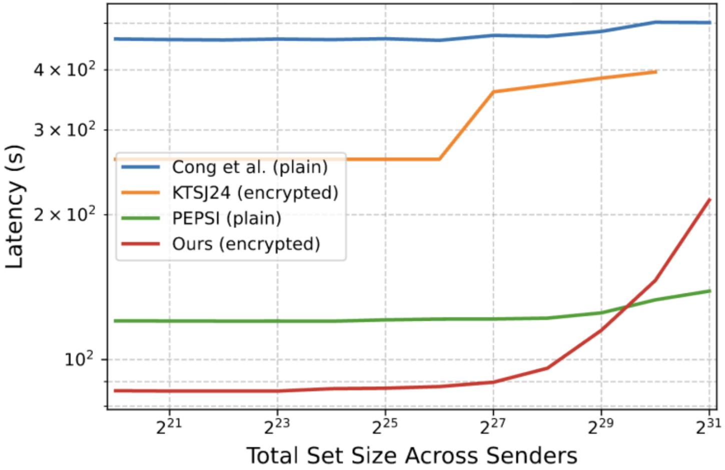




- Selection scales sub-linearly with senders and stays flat across set sizes up to 2^{27}
- At 20 Gbps LAN, 4K senders / 2^{30} items complete in < 400s / < 150s respectively

Selection Stage Comparison





vs. Our Protocol	Speedup	Our Advantage
Cong et al. (plaintext sets)	**3.5× – 6.8×**	We operate on fully encrypted sets
KTSJ24 (encrypted sets)	**3.5× – 6.8×**	Lower depth, higher precision, larger δ
PEPSI (plaintext sets)	**1.4× – 5.4×**	PEPSI uses plaintext; ours is faster at > 4K senders

Performance on Real-World Fraud Datasets

Fraud Dataset Benchmark (FDB) · Downstream: logistic regression · δ = 64, κ = 8

VLDP

Records	233K
Label types	44
Cloud Servers	176

62.6 s

CCTFD

Records	1.2M
Label types	24
Cloud Servers	960

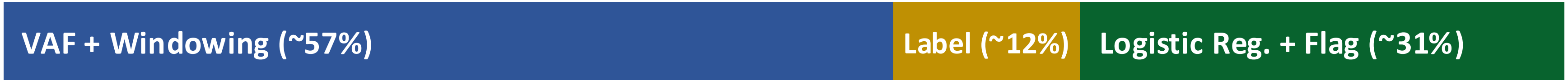
63.5 s

IEEE-CIS

Records	590K
Label types	25
Cloud Servers	250

58.9 s

Runtime Breakdown (8 threads)



VAF + Windowing (~57%)

Label (~12%)

Logistic Reg. + Flag (~31%)

All datasets complete end-to-end in under 65 seconds. Extending to δ = 128 adds only ~11–41% latency

Results and Takeaways

Select + Compute on Encrypted Labels (ELSA)

First CKKS-based protocol for encrypted label selection and real-valued downstream analytics

Novel VAF with Provable Accuracy

wDEP + Bell-shaped composition + Slotwise-windowing supports 2^{64}–2^{128} domains w/ efficiency

Practical and Scalable

Under 65 sec on real-world fraud datasets; scales to 1000s of senders and 2^{30} items

Up to 6.8× Faster

Speedup over state-of-the-art, while operating on fully encrypted datasets

Future Work: Multi-query workloads · Richer predicates · System-level optimizations

Thank you

Any questions?
Email: nkoirala@nd.edu
Website: n7koirala.github.io

Link to the paper

Link to the code

UNIVERSITY OF
NOTRE DAME