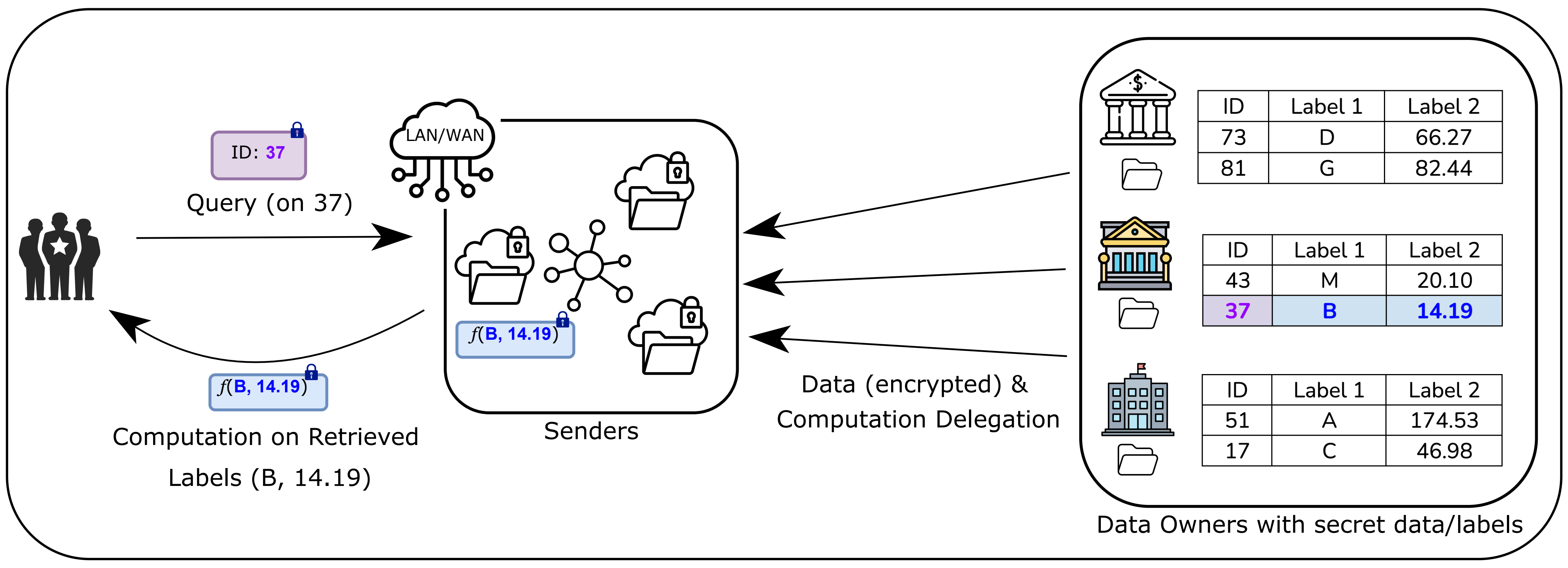
Select-Then-Compute: Encrypted Label Selection and Analytics over Distributed Datasets using FHE

Nirajan Koirala*, Seunghun Paik[†], Sam Martin*, Helena Berens*, Tasha Januszewicz*, Jonathan Takeshita[‡], Jae Hong Seo[†], Taeho Jung*
*University of Notre Dame, [†]Hanyang University, [‡]Old Dominion University

Abstract

Private Set Intersection (PSI) protocols allow a querier to determine whether an item exists in a dataset without revealing the query. However, existing protocols cannot perform downstream computations on associated labels after intersection. We present **ELSA**, the first *encrypted label selection and analytics* protocol that allows secure retrieval of matched labels and downstream function evaluation (e.g., ML inference) on encrypted, distributed datasets using CKKS-based FHE. Our protocol achieves **1.4× to 6.8× speedup** over prior approaches and processes encrypted labels in **under 65 seconds**.

ELSA Protocol Overview


ID: 37
Query (on 37)
LAN/WAN
f(B, 14.19)
f(B, 14.19)
Computation on Retrieved Labels (B, 14.19)
Senders
Data (encrypted) & Computation Delegation
Data Owners with secret data/labels

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

1 Background

- Modern collaborative workflows in finance and healthcare often require identifying common records across independent custodians while strictly adhering to privacy regulations.
- Private Set Intersection (PSI) protocols allow finding matches, but traditional variants cannot securely compute on the associated sensitive labels (payloads) without revealing them to the querier.
- Fully Homomorphic Encryption (FHE), specifically the CKKS scheme, enables approximate arithmetic on encrypted real-valued data, which is essential for machine learning and statistical analytics.

2 Real-World Applications

- **Anti-Money Laundering (AML):** Banks compute risk scores across institutions without revealing customer data
- **Healthcare Analytics:** Privacy-preserving disease risk prediction across distributed patient records
- **Watchlist Screening:** Secure queries against encrypted regulatory databases
- **Fraud Detection:** Cross-institutional fraud scoring on encrypted transaction data

3 Motivation

- Real-world applications, such as anti-money laundering (AML) and healthcare analytics, require not just matching identifiers but also performing complex floating-point computations on their labels.
- Existing methods fail to meet these needs: Labeled-PSI exposes labels, while Circuit-PSI struggles with complex real-valued functions and needs expensive MPC for scaling to large, distributed environments.
- High-precision approximation for large identifier domains (e.g., 64 to 128 bits) is computationally prohibitive in standard FHE, creating a bottleneck for practical, large-scale deployment.

Key Challenge: Execute both *label selection* and *label analytics* while keeping all data encrypted end-to-end with low communication/computation.

4 Contributions

1. We propose the ELSA protocol, the first CKKS-based protocol that integrates secure label selection with real-valued downstream analytics on distributed datasets.
2. We introduce two novel techniques: (a) novel approximation for homomorphic equality testing using wDEPs and Bell-shaped functions, and (b) slot-wise windowing, to efficiently handle large identifiers (2^{64} to 2^{128}) with low FHE depth.
3. We achieve up to 6.8× speedup over state-of-the-art, scaling to thousands of senders and processing real-world fraud datasets in under 65 seconds.

5 Protocol Design

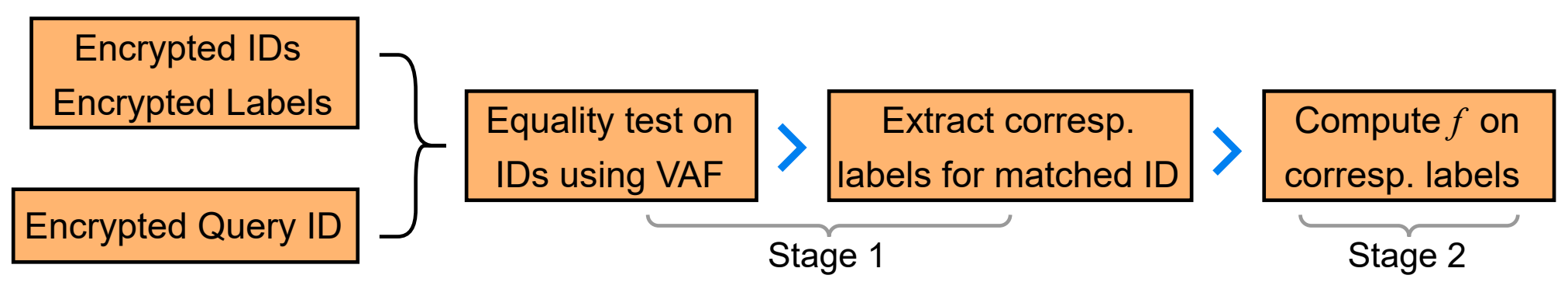

Encrypted IDs
Encrypted Labels
Encrypted Query ID
Equality test on IDs using VAF
Extract corresp. labels for matched ID
Compute f on corresp. labels
Stage 1
Stage 2

Figure 1: Core components of the ELSA protocol

Two-Stage Pipeline

Stage 1 – Label Selection & Extraction:

1. Each sender computes encrypted difference between identifiers using SIMD in FHE
2. Apply Value Annihilating Function (homomorphic equality) using wDEPs and Bell-shaped functions to obtain an indicator ciphertext
3. Multiply indicator with label ciphertext to extract matching label ciphertext
4. Transmit matching label ciphertext to leader sender

Stage 2 – Computation on Labels:

1. Leader aggregates all label ciphertexts
2. Homomorphically evaluate function $f(\text{labels}; \Theta)$
3. Return only the encrypted final result to the receiver

6 Novel VAF Construction

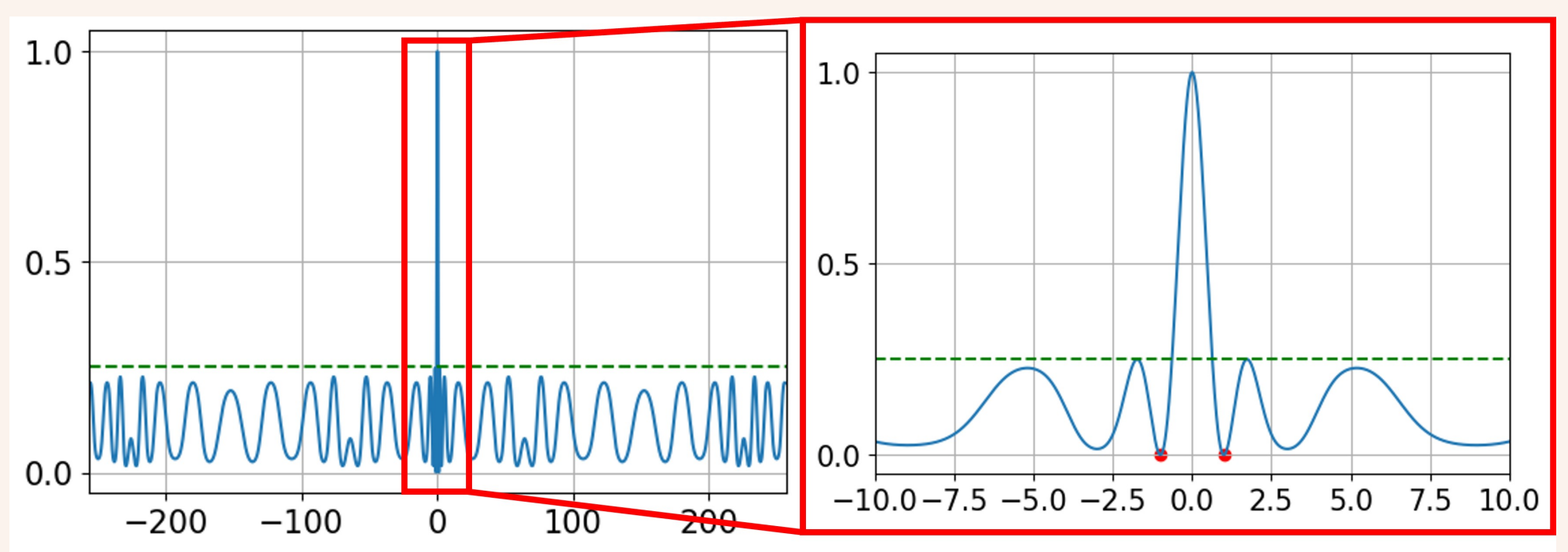


Figure 2: Final VAF from composition of wDEPs and Bell-shaped functions

- **Weak DEPs (wDEPs):** Compress wide input ranges without endpoint saturation
- **Bell-shaped functions:** Concentrate mass at zero with provable bounds. We prove closed-form bounds on approximation error and depth

7 Slot-Wise Windowing

For large identifier domains ($\delta = 64$ or 128 bits), we split each identifier into κ smaller chunks:

$$f_{\text{VAF}}(x) := \prod_{i=1}^{\kappa} f_{\text{VAF},\xi}(x_i)$$

- Reduces domain from 2^{64} to 2^{8} ($\kappa = 8$)
- Parallel VAF computation per chunk
- Only $\log_2 \kappa$ depth for final product
- False positive rate $< 2^{-100}$ for $\delta = 128$

8 Evaluation Results

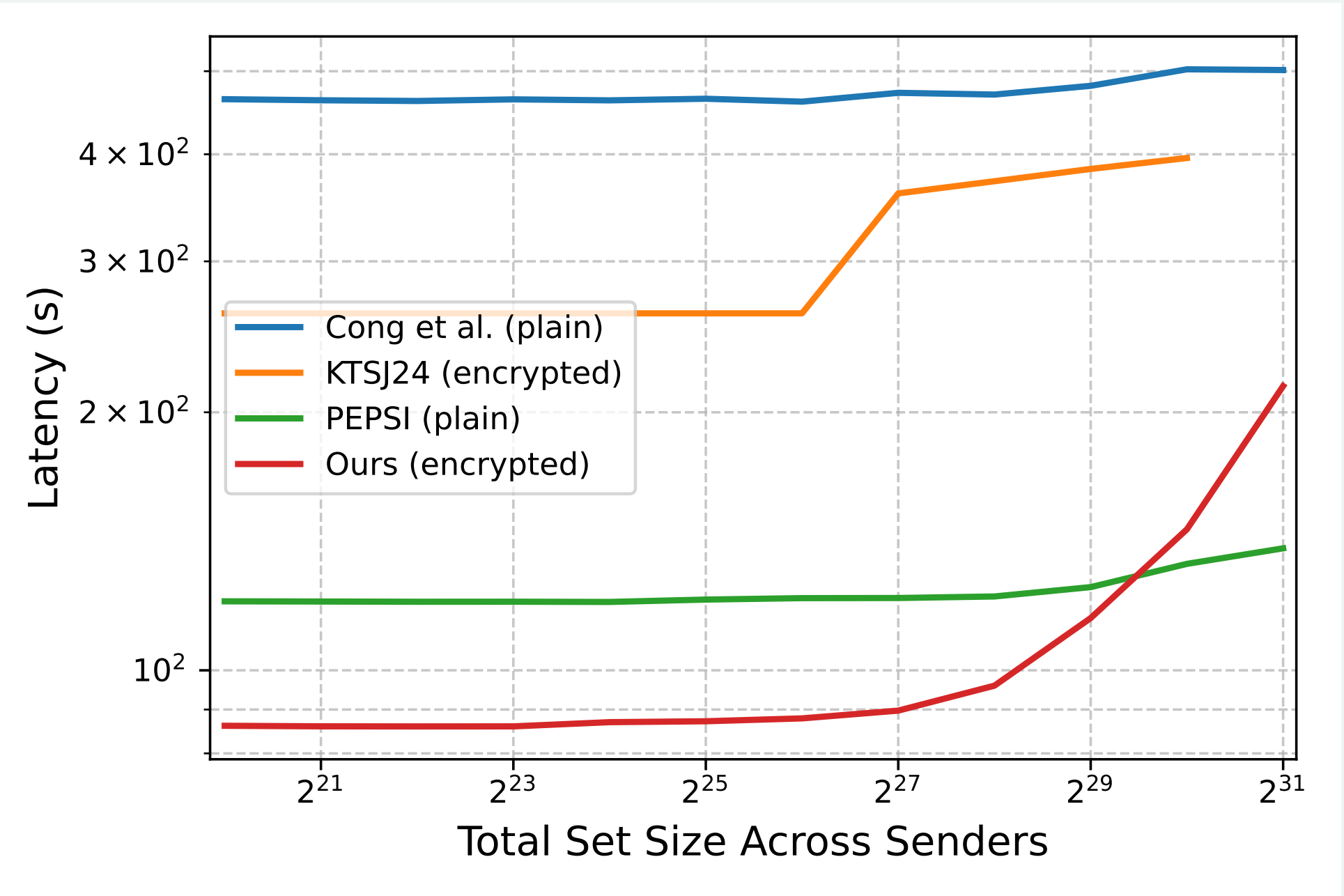


Figure 3: Runtime comparison: **1.4×–5.4× speedup** over PEPSI (Mahdavi et al., USENIX Security 2024), KTSJ24 (Koirala et al., PoPETS 2024), and Cong et al. (ACM CCS 2021) ($\delta = 64$, $\kappa = 8$, 1024 senders).

Dataset	Entries	Senders	Latency
Vehicle Loan	233K	176	62.6 sec
IEEE-CIS Fraud	590K	250	58.9 sec
CCFDT (Credit Card)	1.2M	960	63.5 sec

Evaluation Results for Real-World Datasets

- End-to-end logistic regression on encrypted labels
- VAF selection dominates runtime (>50%)
- 128-bit items add only ~15% overhead
- Matches plaintext-level accuracy up to 24 bits

9 Security Model

- **Threat Model:** Semi-honest adversary corrupting up to $n/2$ parties (honest majority)
- **Security:** IND-CPA^{D} via threshold CKKS FHE with noise flooding via static noise estimation
- **Privacy Guarantees:**
  - ○ Query is hidden from all senders
  - ○ Labels ciphertext is hidden from receiver (only f(labels) revealed)
  - ○ Non-matching records are never exposed

10 Conclusions

- **First protocol** for encrypted label selection & analytics with real-valued computations
- **Novel VAF + slot-wise windowing** enable equality testing over 64/128-bit domains
- Up to **6.8× speedup** over state-of-the-art
- **Under 65 sec** end-to-end latency on real-world fraud datasets

Acknowledgments: This research was supported by NSF CNS-2337321, OAC-2312973, and the Korea Creative Content Agency (RS-2024-00332210).

Code: `https://doi.org/10.5281/zenodo.17849201`

Network and Distributed System Security (NDSS) Symposium 2026

Contact: {nkoirala, tjung}@nd.edu