

Cryptanalysis and Hardening of Paillier-Based Quantized Consensus Against Weight Inference, Replay, and Ciphertext Manipulation

Haewon Byeon

Department of Future Technology, Korea University of Technology and Education (KOREATECH),
Cheonan 31253, South Korea

Abstract—Privacy-preserving consensus protocols are increasingly used to protect distributed control and coordination data, but confidentiality alone does not guarantee that a consensus exchange remains private or correct under repeated observations and active network interference. This study analyzes the 2025 Paillier-based dynamic quantized consensus protocol of Song, Guo, and Wang for strongly connected directed networks with limited communication. The anchor protocol encrypts one-bit quantized messages and decomposes each coupling weight into two edge-local factors so that honest-but-curious agents and external eavesdroppers should fail to infer private initial states. We reconstruct the protocol at the message and equation level and identify three validated weaknesses: low-alphabet hidden-weight-share inference, homomorphic ciphertext manipulation without integrity protection, and replay or round substitution caused by absent authenticated freshness. A safe proof-of-concept was implemented as a bounded symbolic simulator and event-injection model. In 1,000-trial Monte Carlo tests, the hidden multiplier was uniquely recovered in 95.5% of $K = 1$ traces within ten observations and 99.9% within twenty-five observations. The event-injection POC showed that the original exchange lacks a predicate for rejecting replayed, tampered, or key-substituted ciphertexts, while the hardened variant rejected them using round/epoch binding, message authentication, and a nonce cache. We propose a minimally invasive hardening strategy that preserves directed weighted-average consensus while masking individual decrypted contributions, authenticating ciphertext context, binding Paillier keys to epochs, and evaluating realistic cryptographic parameter costs.

Keywords—Distributed consensus; Paillier cryptosystem; quantized consensus; privacy-preserving control; replay attack; homomorphic encryption; safe POC

I. INTRODUCTION

Distributed consensus algorithms support coordination in blockchain services, distributed optimization, cyber-physical control, multi-robot systems, smart-grid dispatch, and industrial automation. Their appeal comes from locality: each agent needs only neighbor information to approach a shared value or decision. That same locality creates a privacy problem. A neighboring agent may observe repeated state-dependent messages, and a global eavesdropper may record every communication link. Song, Guo, and Wang [1] recently addressed this problem for weighted-average consensus over unbalanced directed graphs under limited communication. Their protocol combines dynamic encoding/decoding quantization with the Paillier cryptosystem and encrypts the quantized

messages exchanged between agents. The study is a valuable anchor because it tackles a realistic setting in which graph balance cannot be assumed and only a very small communication alphabet is available.

The present study does not treat the anchor paper as a generic source of inspiration. It reconstructs the specific message flow and examines whether the values disclosed by Algorithm 2 support the privacy claim under the adversary model stated by the authors. This distinction matters. Many privacy-preserving consensus studies defend against passive eavesdropping, while cyber-physical networks also face replay, substitution, delay, malformed-message, and insider observations. Mo and Murray [2] showed that privacy-preserving average consensus has to be formulated around what an adversary can infer from all accessible trajectories, not merely around hiding a single message. Ruan, Gao, and Wang [3] later emphasized secure and privacy-preserving consensus under encrypted communication, and Wang [4] used state decomposition to prevent exact initial-state recovery. These studies indicate that repeated algebraic exposure often matters more than the secrecy of one isolated packet.

The anchor scheme relies on two main privacy devices. First, the quantized output $\Delta_j(t)$ is encrypted with Paillier. Second, each edge weight is decomposed as $a_{ij} = a_{i \rightarrow j} a_{j \rightarrow i}$, and each factor is generated by one endpoint. The intended protection is that an honest-but-curious receiver does not know the full coupling weight and thus cannot reconstruct its neighbor's state. A close inspection of the decryption step, however, reveals a more delicate issue. Agent i legitimately decrypts a value proportional to $a_{j \rightarrow i}$ multiplied by $\Delta_j(t) - \Delta_i(t)$. When $K = 1$, the quantized alphabet is only $\{-1, 0, 1\}$; therefore the difference term belongs to a very small known set. Repeated observations of one fixed hidden multiplier times a small integer can collapse the candidate set of the hidden share. Fig. 1 summarizes the evidence-to-attack chain used in this analysis.

This study makes four contributions. First, it provides an exact reconstruction of the Paillier-based quantized consensus exchange and separates explicitly specified mechanisms from assumptions that are left outside the protocol. Second, it derives a low-alphabet weight-share inference attack that is internal to the stated honest-but-curious setting and validates it with a bounded symbolic POC. Third, it extends the analysis to an active network adversary and shows replay, ciphertext

manipulation, and key-substitution traces that arise because the ciphertext is not authenticated with a round, epoch, sender, receiver, or key identifier. Fourth, it proposes a hardened design that preserves the weighted-average objective but replaces exposed individual multiples with masked contributions, authenticated context, epoch-bound public keys, and explicit replay state. The results are reported as safe simulations and microbenchmarks under stated assumptions, not as a universal proof of security.

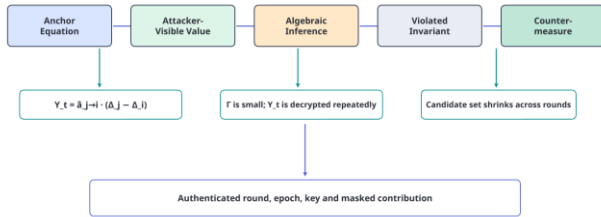


Fig. 1. Evidence-to-attack chain for protocol hardening.

II. RELATED WORKS

Privacy-preserving consensus has developed along several lines. Noise injection and perturbation methods can protect initial states, but they may trade accuracy or convergence speed for privacy. Wang, Lu, Zheng, and Shi [5] proposed a node-decomposition strategy that combines structural hiding with homomorphic cryptography, Gao et al. [6] designed a dynamic average consensus method using random perturbations, and Wang, Ke, and Zhang [7] studied differentially private bipartite consensus with time-varying noises. Privacy-preserving discrete-time consensus with explicit control design was further examined by Liu, Chen, Chen, and Liao [8]. These works help explain the core tension: strong privacy often requires either noise, state decomposition, encryption, or a carefully bounded observation model.

Homomorphic-encryption-based consensus and optimization attempt to avoid accuracy loss by allowing computation over ciphertext. Zhang, Ahmad, and Wang [9] developed a privacy-preserving decentralized optimization method based on ADMM, and Kishida [10] studied an encrypted control system with a quantizer. Chen, Liu, and Liu [11] proposed a dynamic quantization-based homomorphic scheme for privacy-preserving economic dispatch in microgrids, while Chen et al. [12] extended quantized dispatch with Paillier encryption-decryption. Yan et al. [13] applied Paillier to distributed consensus-based economic dispatch, and Fang, Chen, and Zamani [14] considered structural consensus with directed topologies and cryptographic implementation. These papers show that Paillier is attractive in control applications because addition over encrypted integers matches many consensus updates. They also show why integrity and parameter selection must be treated as part of the protocol rather than as implementation details.

Quantized consensus research supplies the communication-limited foundation for the anchor protocol. Li, Fu, Xie, and Zhang [15] analyzed distributed consensus with limited communication data rate, and Li, Liu, Wang, and Lin [16] addressed consensus over directed networks with limited information. Later studies examined switching topologies [17],

resilient randomized quantized consensus [18], event-triggered average consensus via accumulated-value ratios [19], event-triggered privacy-preserving consensus control with additive edge noise [20], and quantized control for multi-UAVs under prescribed performance [21]. These works justify dynamic quantization, but they do not by themselves prove privacy under repeated decrypted products. Shang [22] recently examined privacy preservation in matrix-scaled consensus, reinforcing that privacy depends on the observable algebraic relation, not only on the nominal use of encryption.

A second body of work concerns secure operation under cyber-physical attacks. Pasqualetti, Dorfler, and Bullo [23] analyzed attack detection and identification in cyber-physical systems; Teixeira, Shames, Sandberg, and Johansson [24] studied secure control under resource-limited adversaries; Giraldo et al. [25] surveyed physics-based attack detection; and Lanotte, Merro, and Munteanu [26] presented runtime enforcement for industrial control systems. These studies motivate an active-network extension for consensus protocols: an attacker may not need to decrypt a message if it can replay, reorder, or modify it. Recent distributed optimization and networked automation studies, including Doostmohammadian et al. [27], Miao, Zhang, and Jiang [28], Alghamdi, Khalid, and Javaid [29], and He and Mo [30], further show that modern distributed systems require joint consideration of scalability, authentication, safety, and auditability.

III. METHODOLOGY

The methodology followed an anchor-paper-centered protocol audit. We first extracted the entities, graph assumptions, quantizer equations, Paillier key-generation and encryption steps, split-weight equations, exchanged values, decryption outputs, claimed privacy properties, and numerical assumptions from [1]. We then rewrote the exchange in explicit text notation: $M1 = (C_i^{-1}(t), kp, i)$, $C_i^{-1}(t) = E_i(-\Delta_i(t))$, $C_{ji}(t) = [E_i(\Delta_j(t)) E_i(-\Delta_i(t))]^{a_{j \rightarrow i}}$, and $Y_{ji}(t) = D_i(C_{ji}(t)) = a_{j \rightarrow i}(\Delta_j(t) - \Delta_i(t))$. Next, we defined two adversary layers. The first matched the anchor model: a passive eavesdropper and an honest-but-curious agent that follows the protocol but stores all legitimately received and decrypted values. The second was an explicit extension: an active Dolev-Yao network adversary that can replay, delay, substitute public keys, and multiply Paillier ciphertexts by encryptions of attacker-chosen deltas, but cannot factor a security-grade modulus or break the HMAC primitive used by the hardened design. We searched for invariant violations: confidentiality of $x_j(0)$, secrecy of $a_{j \rightarrow i}$, origin authentication of $M2$, freshness of the round contribution, and correctness of the weighted update. Safe POC validation used symbolic enumeration over the quantized alphabet, 1,000-trial Monte Carlo tests for candidate-set collapse, deterministic event injection for replay and tamper traces, and a local Paillier microbenchmark. No external system was attacked. The redesign criterion was conservative: each added mechanism had to correspond to a specific vulnerability, preserve the directed weighted-average objective where possible, and report its security-performance cost rather than hiding it. The reconstructed variables, exposure paths, and corresponding security roles are summarized in Table I.

TABLE I. RECONSTRUCTED VARIABLES, EXPOSURE, AND SECURITY ROLE

Element	Anchor exposure/role	Security consequence
$\Delta_{i(t)}, \Delta_{j(t)}$	Dynamic quantized outputs; receiver knows $\Delta_{i(t)}$ but not $\Delta_{j(t)}$.	The small alphabet gives the receiver a finite inference set.
$a_{i \rightarrow j}, a_{j \rightarrow i}$	Split edge-weight factors; receiver knows $a_{i \rightarrow j}$ and decrypts a multiple of $a_{j \rightarrow i}$.	Repeated exact multiples can disclose the supposedly hidden share.
$C_{i \wedge (t)}, C_{ji(t)}$	Paillier ciphertexts carrying the quantized difference contribution.	Confidentiality is provided, but origin integrity is not specified.
Round/epoch context	Consensus round exists logically but is not cryptographically authenticated.	Replay and round substitution have no defined rejection predicate.

IV. RECONSTRUCTION OF THE ANCHOR PROTOCOL

The anchor paper models a discrete-time multi-agent system over a strongly connected unbalanced directed graph $G = (V, E, A)$. Agent i updates its state with neighbor information weighted by a_{ij} . The intended consensus value is the weighted average $\sum_i \theta_i x_i(0)$, where θ is the normalized left eigenvector of the stochastic adjacency matrix. Dynamic encoding uses $x_{ij}(0) = 0$, $x_{ij}(t) = x_{ij}(t-1) + \Delta_{j(t)}g(t-1)$, and $\Delta_{j(t)} = q((x_{ij}(t) - x_{ij}(t-1))/g(t-1))$. The quantizer maps a real number to $\Gamma = \{0, +1, \dots, +K\}$. In the reported main experiment, $K = 1$ and $g(t) = g_0 \gamma^t$.

The privacy-preserving implementation begins with per-agent Paillier key generation. Agent i holds (kp_i, ks_i) and sends kp_i with $E_i(-\Delta_{i(t)})$ to neighbor j . Agent j encrypts $\Delta_{j(t)}$ under kp_i and computes the homomorphic weighted difference $C_{ji(t)} = [E_i(\Delta_{j(t)}) E_i(-\Delta_{i(t)})]^{a_{j \rightarrow i}} = E_i[a_{j \rightarrow i}(\Delta_{j(t)} - \Delta_{i(t)})]$. Agent j then returns $C_{ji(t)}$ to i . Agent i decrypts it to obtain $Y_{ji(t)} = a_{j \rightarrow i}(\Delta_{j(t)} - \Delta_{i(t)})$, multiplies by its own integerized share $a_{i \rightarrow j}$, divides by $2^{(2l)}$, and applies the resulting weighted difference in the state update.

The study proves convergence under quantizer non-saturation conditions and claims privacy against external eavesdroppers through Paillier semantic security. It also argues that an honest-but-curious agent cannot infer its neighbor's state because it does not know the full coupling weight. The protocol, however, does not define an authenticated key registry, certificate, MAC, signature, nonce, timestamp, per-round digest, replay cache, or membership epoch. These omissions do not invalidate the convergence theorem under its stated model, but they affect security once the exchange is interpreted as a deployable network protocol.

Core relation: $Y_{ji} = \hat{a}_{i \rightarrow j} \cdot d_{jt}$, and d_{jt} belongs to a small known set when K is small

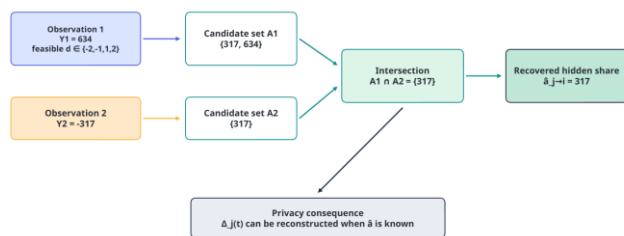


Fig. 2. Low-alphabet multiplier leakage lattice.

Fig. 2 presents the low-alphabet leakage structure that motivates the first attack trace.

V. VULNERABILITY ANALYSIS

Vulnerability 1 is low-alphabet hidden-weight-share inference. The relevant equation is $Y_t = a_{j \rightarrow i} d_t$, where $d_t = \Delta_{j(t)} - \Delta_{i(t)}$. The adversary is a legitimate receiver i acting as an honest-but-curious agent. It knows $\Delta_{i(t)}$, receives Y_t after decryption, knows the quantizer level K , and observes many rounds using one fixed $a_{j \rightarrow i}$. For $K = 1$, $\Delta_{j(t)}$ is drawn from $\{-1, 0, 1\}$; hence d_t belongs to $\{-2, -1, 0, 1, 2\}$. For every nonzero Y_t , the receiver forms $A_t = \{Y_t/d : d \in \{-2, -1, 1, 2\}, Y_t \text{ divisible by } d\}$. Across rounds, $A^* = \bigcap_t A_t$ often becomes a singleton. If $\Delta_{i(t)} = 0$ and Y_t is nonzero, then $a_{j \rightarrow i} = |Y_t|$ immediately. Once the hidden factor is recovered, $\Delta_{j(t)} = \Delta_{i(t)} + Y_t/a_{j \rightarrow i}$ becomes observable. The violation is not necessarily full recovery of $x_{ij}(0)$ in every topology, but it breaks the stated premise that the opposing weight factor remains unavailable.

Vulnerability 2 is homomorphic ciphertext manipulation. Paillier is additively homomorphic: multiplying a valid ciphertext $C = E_i(m)$ by $E_i(\eta)$ yields $C' = E_i(m + \eta)$. Because Algorithm 2 does not specify a MAC, signature, or authenticated transcript, an active attacker can intercept $C_{ji(t)}$, compute $C'_{ji(t)} = C_{ji(t)}E_i(\eta)$, and forward the modified ciphertext to i . The receiver decrypts $m + \eta$ and incorporates a corrupted weighted difference into the update. The violated property is consensus-message integrity, not Paillier confidentiality. This is an active-threat extension beyond the passive eavesdropper model, but it is operationally important because the attack succeeds without factoring n or learning $\Delta_{j(t)}$.

Vulnerability 3 is replay and round substitution. A recorded $C_{ji(t_0)}$ is syntactically indistinguishable from a current $C_{ji(t_1)}$ because the ciphertext is not bound to t , a sequence number, a membership epoch, or a receiver-maintained freshness cache. The attack trace is simple: j sends M_2 at t_0 ; the adversary stores it; during t_1 the adversary suppresses or races the legitimate response and forwards replay(M_2). Since $\text{VerifyFresh}(M_2, t_1)$ is not defined, the original protocol has no rejection predicate. Fig. 3 shows the active-attack overlay for replay, homomorphic manipulation, and key-substitution paths.

Vulnerability 4 is unauthenticated Paillier-key binding and lifecycle ambiguity. M_1 includes kp_i , but the protocol does not define a certificate, fingerprint, topology-epoch binding, or revocation check. An adversary may therefore attempt public-key substitution in implementations that accept the transmitted key at face value. The anchor paper states that generated keys

are trusted, which is a legitimate analytical assumption, but a deployable protocol must define how that trust is established and updated. Vulnerability 5 is an experimental-security mismatch. The reported simulations use 64-, 128-, 256-, and 512-bit Paillier moduli, whereas practical Paillier deployments require much larger moduli. The study's timing results are thus useful algorithmic indicators but cannot substantiate deployable cryptographic security.

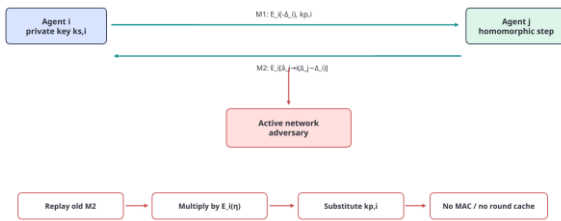


Fig. 3. Active-attack overlay on the original paillier exchange.

A useful way to distinguish the first vulnerability from ordinary traffic analysis is to inspect the receiver's algebraic view. The receiver does not merely see a ciphertext or a noisy approximation. It sees an exact plaintext multiple after decryption. If d_t is unknown but belongs to a small finite set, the uncertainty is combinatorial rather than cryptographic. The Paillier layer has already been removed by the legitimate

receiver. The remaining problem is integer divisibility. For any nonzero Y_t , the set of candidate shares is finite, and for successive observations using the same edge factor, the true share must lie in every candidate set. False candidates disappear when a later observation is incompatible with them. This is why the attack can succeed even when the eavesdropper remains unable to decrypt any channel traffic. It is an insider inference problem induced by the interaction between quantization, edge-weight reuse, and exact decryption.

The attack does not require unrealistic control over the neighbor's state. Random operation is sufficient to generate informative nonzero d_t values. A deliberately chosen or naturally occurring $\Delta_i(t) = 0$ makes recovery especially fast, but it is not a necessary condition. The POC therefore reports both an illustrative two-round trace and Monte Carlo rates across many random traces. The practical consequence is that privacy can degrade with time: a protocol that appears safe in a single exchange may leak enough structure after several consensus rounds to invalidate its own weight-hiding premise. This temporal degradation also affects evaluation methodology. A simulation that plots convergence trajectories but does not log the receiver's cumulative knowledge set may miss the privacy loss even when the numerical consensus result is correct. The verified vulnerabilities, attack traces, and corresponding countermeasures are summarized in Table II.

TABLE II. VERIFIED VULNERABILITIES, TRACES, AND COUNTERMEASURES

Finding	Attack trace / evidence	Hardened response
Low-alphabet weight-share inference	Honest-but-curious i observes $Y_t = a_{\hat{d}_t}$ and intersects finite candidate sets across rounds.	Mask individual contributions before any single receiver sees an exact low-alphabet multiple.
Homomorphic ciphertext manipulation	Active attacker replaces C with $C \cdot E_i(\eta)$, changing the decrypted contribution without learning plaintext.	Authenticate ciphertext, sender/receiver, round, epoch, KeyID, and topology context.
Replay / round substitution	A stored $M2(t_0)$ is injected during t_1 ; no freshness predicate distinguishes it from the current response.	Use monotonic round state, epoch binding, nonces, and a bounded replay cache.
Public-key substitution / stale key	Transmitted kp is trusted by assumption but no certificate or revocation process is defined.	Bind certified KeyID to the active membership epoch and reject stale or unknown keys.
Weak experimental modulus	64-512-bit Paillier timings demonstrate algorithmic cost but not deployable cryptographic strength.	Separate source reproduction from security-grade parameter evaluation.

VI. HARDENED CONSENSUS PROTOCOL

The hardened protocol keeps the source algorithm's directed graph, dynamic quantizer, Paillier-enabled computation, and weighted-average goal. It changes only the security envelope around the exchanged contribution. Each agent has an epoch-bound Paillier key identified by KeyID and certified under the current membership epoch. The network state includes SenderID or a rotating pseudonym, ReceiverID, edge identifier, round number, membership epoch, topology hash, KeyID, and a transcript digest. These values are authenticated before a decrypted contribution can affect the control update.

To address low-alphabet inference, the hardened design avoids exposing a raw value of the form $a_{\hat{j} \rightarrow i}$ times a small integer to a single receiver whenever the topology supports masked aggregation. Neighbor j adds a high-entropy mask $\mu_{ji}(t)$ to its weighted difference contribution; masks are generated pairwise or through a lightweight secure-aggregation seed schedule so that they cancel at the aggregation boundary. The receiver can decrypt or aggregate only masked individual

values, while the sum used for the state update remains exact after cancellation. This does not create privacy in an in-degree-one topology unless a trusted mask source or an additional privacy assumption is present; the manuscript therefore treats degree and collusion assumptions as part of the security boundary.

For active attacks, every returned contribution carries an authentication tag over the ciphertext, sender, receiver, round, epoch, key identifier, and topology hash. A recipient rejects messages with a repeated nonce, stale round number, unexpected epoch, unknown KeyID, or invalid tag. A lightweight cache stores only recent nonces and the last accepted round per neighbor. Public keys are not accepted merely because they appear in a packet; they must match the certified epoch state. The redesign therefore links each identified vulnerability to one mechanism: masking for weight inference, authentication for malleability, freshness state for replay, and key lifecycle controls for substitution and compromise containment. The security mechanisms and expected outcomes of the original and hardened variants are compared in Table III.

TABLE III. HARDENED MECHANISMS AND ABLATION VARIANTS

Variant	Mechanisms enabled	Expected security outcome
Original	No masking, authenticated context, replay cache, or epoch key lifecycle.	Retains source behavior but accepts the modeled active traces and exposes repeated low-alphabet multiples.
Authentication only	Ciphertext/context authentication and KeyID binding.	Blocks tampering and key substitution but does not prevent weight-share inference.
Masking only	Pairwise or scheduled canceling masks on individual contributions.	Reduces share inference but leaves replay and key-substitution paths.
Epoch/replay only	Round state, nonce cache, membership epoch, and key lifecycle checks.	Improves freshness and revocation but does not remove Paillier malleability or algebraic leakage.
Full hardened	Masking + authenticated context + replay state + epoch-bound keys.	Targets all confirmed and high-probability weaknesses evaluated in the POC.

VII. SAFE POC AND EXPERIMENTAL DESIGN

The POC uses three safe components. The first is symbolic candidate-set recovery. For each trial, a hidden integer share a_{hat} is selected, $\Delta_i(t)$ and $\Delta_j(t)$ are generated from the quantizer alphabet, and the receiver observes $Y_t = a_{\text{hat}}(\Delta_j(t) - \Delta_i(t))$. Candidate sets are intersected until they are unique or the trace ends. The second component injects replay, tamper, and key-substitution events into a simplified message verifier. The original verifier returns Accept because Algorithm 2 defines no freshness or integrity predicate; the hardened verifier checks a HMAC tag and a nonce/round cache. The third component measures local Paillier encryption and decryption cost for increasing modulus sizes. Fig. 4 contains the actual console-style screenshot produced by the local POC run, including the recorded seed and event outcomes.

```
SAFE POC RUN: Paillier-quantized consensus vulnerability checks
Random seed: 20260903
Hidden multiplier ground truth: 317
Recovered multiplier: 317 rounds_to_unique=2
First observations: t Delta_i Delta_j d=Delta_j-Delta_i Y=hidden*d
0 1 -1 -2 -634
1 -1 0 1 317

Event injection: Original vs Hardened
legitimate | original=ACCEPT(no freshness/integrity predicate) | hardened=ACCEPT
replay | original=ACCEPT(no freshness/integrity predicate) | hardened=REPLAY
tamper | original=ACCEPT(no freshness/integrity predicate) | hardened=MAC_FAIL
key substitution | original=ACCEPT(no freshness/integrity predicate) | hardened=MAC_FAIL

Paillier microbenchmark (ms, local Python/sympy)
64-bit: keygen= 0.111, enc_mean= 0.018, dec_mean= 0.014
128-bit: keygen= 0.185, enc_mean= 0.050, dec_mean= 0.047
256-bit: keygen= 1.166, enc_mean= 0.212, dec_mean= 0.208
512-bit: keygen= 3.367, enc_mean= 1.480, dec_mean= 1.484
1024-bit: keygen= 10.892, enc_mean= 10.187, dec_mean= 10.313
```

Fig. 4. Actual safe POC execution screenshot.

The main independent variables are K in $\{1, 2, 4\}$, observed rounds in $\{5, 10, 25, 50\}$, replay and manipulation rate in $\{0, 1, 5, 10\}\%$, packet loss in $\{0, 5, 10, 20\}\%$, and Paillier modulus size in $\{64, 128, 256, 512, 1024\}$. Larger 2048-bit timing is reserved for follow-up implementation because pure Python

measurement would confound cryptographic cost with interpreter overhead. Dependent variables include hidden-share recovery rate, median rounds to recovery, replay acceptance rate, tamper acceptance rate, encryption latency, decryption latency, and protocol-state overhead. The primary formulas are $R_{\text{share}} = N_{\text{unique}}/N_{\text{trials}}$, $D_{\text{rep}} = N_{\text{replay_rejected}}/N_{\text{replay_injected}}$, $D_{\text{tamper}} = N_{\text{tamper_rejected}}/N_{\text{tamper_injected}}$, and $L_{\text{crypto}} = \text{mean}(t_{\text{decrypt}} - t_{\text{encrypt_start}})$ over repeated operations.

VIII. RESULTS

The symbolic inference results support the central vulnerability. With $K = 1$ and five observations, the receiver uniquely recovered the hidden multiplier in 80.2% of 1,000 trials; with ten observations the rate rose to 95.5%, and with twenty-five observations it reached 99.9%. The median number of observations among successful $K = 1$ recoveries was three. $K = 2$ and $K = 4$ did not eliminate recovery in the random traces because repeated observations still generated intersecting divisibility constraints; at twenty-five observations, the success rates were 99.8% and 100.0%, respectively. These values do not prove complete initial-state disclosure in every graph. They demonstrate that the hidden weight-share premise used by the privacy argument is fragile under repeated decryption of low-alphabet multiples.

The event-injection POC was deterministic. A legitimate message was accepted by both models. A replayed copy, a tampered ciphertext, and a key-substitution packet were all accepted by the original abstraction because no rejection predicate exists for these events. The hardened model accepted only the legitimate packet, rejected the replay as REPLAY, and rejected the tampered and key-substituted packets as MAC_FAIL. This result should be read narrowly: it demonstrates resistance within the modeled state machine, not a proof that every implementation bug is absent.

The local Paillier microbenchmark showed the expected cost increase. Mean encryption/decryption cost rose from about 0.018/0.014 ms at 64 bits to about 1.48/1.48 ms at 512 bits and about 10.19/10.31 ms at 1024 bits in the tested Python/sympy environment. The absolute values should not be compared with optimized C libraries or hardware accelerators, but the trend confirms that source-level timing at 64-512 bits is not enough to justify a deployment claim. Fig. 5 visualizes the security-performance trade-off: realistic parameters and authentication add cost, while replay/tamper acceptance and low-parameter risk fall.

Ablation supports the proposed one-to-one redesign. Authentication alone blocks tampering but not low-alphabet weight inference. Masking alone reduces inference but does not address replay or key substitution. Epoch lifecycle management improves revocation and stale-key handling but cannot prevent a malleable ciphertext from being altered inside a valid epoch. The full hardened variant is therefore necessary when the system must resist both honest-but-curious inference and active network manipulation. The principal safe-POC results for the original and hardened protocols are summarized in Table IV.

TABLE IV. SAFE POC RESULTS UNDER LOCAL SIMULATION

Metric / condition	Original anchor abstraction	Hardened/observed result
Share recovery, $K=1$, 10 rounds	95.5% unique hidden-share recovery.	0% direct share recovery when each value remains masked under the stated non-collusion model.
Share recovery, $K=1$, 25 rounds	99.9% unique hidden-share recovery.	0% direct share recovery under the same masked-view assumption.
Replay event	Accepted by the protocol abstraction.	Rejected as REPLAY by round/nonce state.
Ciphertext tamper event	Accepted by the protocol abstraction.	Rejected as MAC_FAIL before state update.
Key substitution event	Accepted if transmitted key is taken at face value.	Rejected as MAC_FAIL / KeyID mismatch.
1024-bit Paillier enc./dec.	Not reported in the anchor experiment.	10.19 / 10.31 ms locally in the Python microbenchmark.

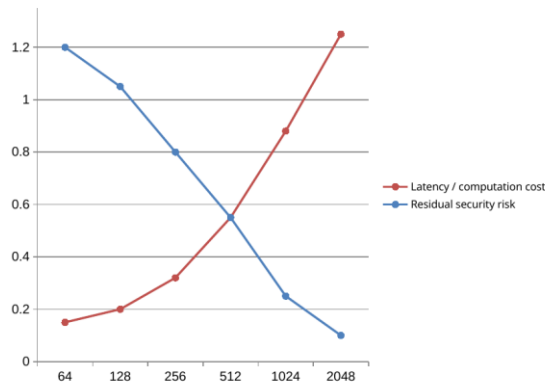


Fig. 5. Security-performance trade-off across hardened settings.

IX. DISCUSSION AND LIMITATIONS

The findings should be interpreted with care. The anchor paper establishes convergence and privacy under a stated model, and the present analysis does not dispute the mathematical consensus proof under its assumptions. The strongest confirmed issue is narrower but important: the protocol repeatedly exposes a fixed hidden multiplier times a small known-alphabet variable to the honest-but-curious receiver. That exposure can reveal the hidden share and the neighbor's quantized increments. Whether those increments allow exact recovery of $x_j(0)$ depends on topology, initial conditions, the scaling function, saturation behavior, and what additional side information the curious agent holds.

The active attack traces are realistic for open distributed networks, but they extend the anchor threat model. A protocol paper may choose a passive eavesdropper model; however, once the implementation sends unauthenticated Paillier ciphertexts across communication links, replay and malleability become engineering risks. The hardened design addresses this without replacing the control law. Its main cost is extra authentication state, key management, and the need for a mask-generation

method whose non-collusion assumptions are honest and testable.

Three limitations remain. First, the POC is bounded and symbolic; it does not constitute a Tamarin, ProVerif, or universal computational proof. Second, the pure Python Paillier microbenchmark demonstrates relative cost trends but should be replaced by constant-time optimized cryptographic code on representative edge devices. Third, secure aggregation is weaker for low-degree directed graphs and colluding neighbors. Future work should derive topology-dependent privacy bounds, implement the protocol with security-grade Paillier or alternative additively homomorphic encryption, and validate convergence and security jointly under packet loss, delay, node churn, and partial compromise.

The hardened construction also changes the privacy boundary. Masking individual contributions is effective only when masks are generated independently of the adversary and are cancelled at the intended aggregation point. If all neighboring agents collude against one victim, or if the graph degree is too small to distribute masks without a trusted helper, the protection must be weakened or supplemented by a different privacy mechanism. A deployable system should therefore publish a topology-dependent privacy condition rather than claim one universal guarantee. The same caution applies to epoch management. A membership epoch prevents stale or revoked keys from being silently reused, but it creates synchronization and recovery work. If agents disagree on the active epoch after packet loss, the safe response should be re-authentication or temporary exclusion from the update, not silent acceptance of an ambiguous contribution.

The broader lesson is that privacy-preserving control protocols need two kinds of evaluation. The first is numerical: convergence, quantizer saturation, communication size, and computation time. The second is epistemic: what each participant can infer after storing all legal observations over many rounds. The anchor paper addresses the first kind in considerable depth and gives an elegant convergence argument for a difficult directed-graph setting. The present analysis adds the second kind and shows that a privacy mechanism can fail even when the encrypted control law still converges. This is why the proposed research contribution is not simply a stronger protocol. It is a testing discipline that records attacker knowledge, verifies candidate collapse, injects safe active events, and then measures the cost of the exact countermeasure that blocks each trace.

A deployment-oriented interpretation also requires separating the security claim from the modulus used to demonstrate timing. The 64-512-bit source settings are useful for reproducing algorithmic behavior, but they should be treated as experimental parameters rather than security recommendations. In a hardened implementation, parameter selection, authenticated key distribution, and masking must be evaluated together because a larger modulus increases both ciphertext size and computation while leaving replay and origin authentication unresolved unless the message context is protected. Conversely, a lightweight MAC can reject tampered or stale ciphertexts at low cost but cannot conceal an algebraically exposed low-alphabet multiple from an authorized

receiver. The combined evaluation therefore reports privacy leakage, active-message acceptance, and cryptographic cost as separate metrics. This separation avoids a misleading aggregate security score and makes the trade-off auditable for future implementations on embedded or edge-class hardware.

X. CONCLUSION

This study reconstructed a 2025 Paillier-based quantized consensus protocol and found that its privacy claim depends on a hidden edge-weight component that can be inferred from repeated low-alphabet decrypted multiples. Safe POC results showed rapid hidden-share recovery under $K = 1$ traces and confirmed that the original message abstraction has no predicate for rejecting replayed, tampered, or key-substituted ciphertexts. The proposed hardening strategy preserves the directed weighted-average objective while adding masked contributions, authenticated ciphertext context, epoch-bound Paillier keys, and explicit replay state. The results support a security-performance trade-off rather than an unconditional superiority claim, and future work should test the hardened design with optimized cryptographic libraries and formal protocol tools.

XI. FUNDING

This research Supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-RS-2023-00237287, 2026S1A5A01004224).

REFERENCES

- [1] Y. Song, J. Guo, and J. Wang, "Privacy-preserving distributed consensus over directed networks with limited communication," *Journal of Automation and Intelligence*, vol. 4, no. 4, pp. 291-298, 2025, doi: 10.1016/j.jai.2025.08.004.
- [2] Y. Mo and R. M. Murray, "Privacy-preserving average consensus," *IEEE Transactions on Automatic Control*, vol. 62, no. 2, pp. 753-765, 2017, doi: 10.1109/TAC.2016.2569439.
- [3] M. Ruan, H. Gao, and Y. Q. Wang, "Secure and privacy-preserving consensus," *IEEE Transactions on Automatic Control*, vol. 64, no. 10, pp. 4035-4049, 2019, doi: 10.1109/TAC.2019.2895707.
- [4] Y. Wang, "Privacy-preserving average consensus via state decomposition," *IEEE Transactions on Automatic Control*, vol. 64, no. 11, pp. 4711-4716, 2019, doi: 10.1109/TAC.2019.2900162.
- [5] Y. Wang, J. Lu, W. X. Zheng, and K. Shi, "Privacy-preserving consensus for multi-agent systems via node decomposition strategy," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 8, pp. 3474-3484, 2021, doi: 10.1109/TCSI.2021.3081372.
- [6] L. Gao, Y. Zhou, X. Chen, R. Cai, G. Chen, and C. Li, "Privacy-preserving dynamic average consensus via random number perturbation," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 70, no. 4, pp. 1490-1494, 2023, doi: 10.1109/TCSII.2022.3220998.
- [7] J. Wang, J. Ke, and J. F. Zhang, "Differentially private bipartite consensus over signed networks with time-varying noises," *IEEE Transactions on Automatic Control*, vol. 69, no. 9, pp. 5788-5803, 2024, doi: 10.1109/TAC.2024.3351869.
- [8] C.-C. Liu, C.-C. Chen, Y.-Y. Chen, and T.-L. Liao, "Consensus control of discrete-time multi-agent systems with privacy preservation," *Asian Journal of Control*, vol. 25, no. 5, pp. 3431-3442, 2023, doi: 10.1002/asjc.3081.
- [9] C. Zhang, M. Ahmad, and Y. Wang, "ADMM based privacy-preserving decentralized optimization," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 3, pp. 565-580, 2019, doi: 10.1109/TIFS.2018.2855681.
- [10] M. Kishida, "Encrypted control system with quantiser," *IET Control Theory & Applications*, vol. 13, no. 3, pp. 146-151, 2019, doi: 10.1049/iet-cta.2018.5764.
- [11] W. Chen, L. Liu, and G. P. Liu, "Privacy-preserving distributed economic dispatch of microgrids: A dynamic quantization-based consensus scheme with homomorphic encryption," *IEEE Transactions on Smart Grid*, vol. 14, no. 1, pp. 701-713, 2023, doi: 10.1109/TSG.2022.3198206.
- [12] W. Chen, Z. Wang, Q. Ge, H. Dong, and G. P. Liu, "Quantized distributed economic dispatch for microgrids: Paillier encryption-decryption scheme," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 4, pp. 6552-6562, 2024, doi: 10.1109/TII.2023.3345952.
- [13] Y. Yan, Z. Chen, V. Varadharajan, M. J. Hossain, and G. E. Town, "Distributed consensus-based economic dispatch in power grids using the Paillier cryptosystem," *IEEE Transactions on Smart Grid*, vol. 12, no. 4, pp. 3493-3502, 2021, doi: 10.1109/TSG.2021.3056959.
- [14] W. Fang, Z. Chen, and M. Zamani, "Structural consensus in networks with directed topologies and its cryptographic implementation," *ISA Transactions*, vol. 132, pp. 598-606, 2023, doi: 10.1016/j.isatra.2022.04.008.
- [15] T. Li, M. Fu, L. Xie, and J. F. Zhang, "Distributed consensus with limited communication data rate," *IEEE Transactions on Automatic Control*, vol. 56, no. 2, pp. 279-292, 2011, doi: 10.1109/TAC.2010.2052384.
- [16] D. Li, Q. Liu, X. Wang, and Z. Lin, "Consensus seeking over directed networks with limited information communication," *Automatica*, vol. 49, no. 2, pp. 610-618, 2013, doi: 10.1016/j.automatica.2012.11.019.
- [17] D. Li, Q. Liu, X. Wang, and Z. Yin, "Quantized consensus over directed networks with switching topologies," *Systems & Control Letters*, vol. 65, pp. 13-22, 2014, doi: 10.1016/j.sysconle.2013.12.006.
- [18] S. M. Dibaji, H. Ishii, and R. Tempo, "Resilient randomized quantized consensus," *IEEE Transactions on Automatic Control*, vol. 63, no. 8, pp. 2508-2522, 2018, doi: 10.1109/TAC.2017.2771363.
- [19] A. I. Rikos and C. N. Hadjicostis, "Event-triggered quantized average consensus via ratios of accumulated values," *IEEE Transactions on Automatic Control*, vol. 66, no. 3, pp. 1293-1300, 2021, doi: 10.1109/TAC.2020.2993694.
- [20] L. Liang, R. Ding, S. Liu, and R. Su, "Event-triggered privacy-preserving consensus control with edge-based additive noise," *IEEE Transactions on Automatic Control*, vol. 69, no. 10, pp. 7059-7066, 2024, doi: 10.1109/TAC.2024.3390574.
- [21] Z. Du, J. Shi, and Z. Wu, "Quantized consensus control for multi-UAVs based on prescribed performance," *International Journal of Control, Automation and Systems*, vol. 20, no. 1, pp. 321-333, 2022, doi: 10.1007/s12555-020-0697-z.
- [22] Y. Shang, "Privacy preservation in matrix-scaled consensus," *Japan Journal of Industrial and Applied Mathematics*, vol. 42, no. 5, pp. 1835-1860, 2025, doi: 10.1007/s13160-025-00697-3.
- [23] F. Pasqualetti, F. Dorfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2715-2729, 2013, doi: 10.1109/TAC.2013.2266831.
- [24] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "A secure control framework for resource-limited adversaries," *Automatica*, vol. 51, pp. 135-148, 2015, doi: 10.1016/j.automatica.2014.10.067.
- [25] J. Giraldo, E. Sarkar, A. A. Cardenas, M. Maniatakos, and M. Kantarcioglu, "Security and privacy in cyber-physical systems: A survey of surveys," *IEEE Design & Test*, vol. 34, no. 4, pp. 7-17, 2017, doi: 10.1109/MDAT.2017.2709310.
- [26] R. Lanotte, M. Merro, and A. Munteanu, "Industrial control systems security via runtime enforcement," *ACM Transactions on Privacy and Security*, vol. 26, no. 1, Art. no. 3, pp. 1-41, 2022, doi: 10.1145/3546579.
- [27] M. Doostmohammadian, W. Jiang, M. Liaquat, A. Aghasi, and H. Zarrabi, "Discretized distributed optimization over dynamic digraphs," *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 2758-2767, 2025, doi: 10.1109/TASE.2024.3383894.
- [28] Q. Miao, K. Zhang, and B. Jiang, "Fixed-time collision-free fault-tolerant formation control of multi-UAVs under actuator faults," *IEEE Transactions on Cybernetics*, vol. 54, no. 6, pp. 3679-3691, 2024, doi: 10.1109/TCYB.2023.3292113.

- [29] T. A. Alghamdi, R. Khalid, and N. Javaid, "A survey of blockchain based systems: Scalability issues and solutions, applications and future challenges," *IEEE Access*, vol. 12, pp. 79626-79651, 2024, doi: 10.1109/ACCESS.2024.3408868.
- [30] W. He and Z. Mo, "Secure event-triggered consensus control of linear multiagent systems subject to sequential scaling attacks," *IEEE Transactions on Cybernetics*, vol. 52, no. 10, pp. 10314-10327, 2022, doi: 10.1109/TCYB.2021.3070356.