POSTER: A Structured Framework for the Reproducible Evaluation of Privacy and Anonymity

Anonymous Author(s)

Abstract

Privacy and anonymity are foundational security properties in modern systems, yet empirical evaluations of these properties often suffer from limited reproducibility and weak comparability across studies. Prior work frequently relies on implicit threat models, underspecified adversary capabilities, and experiment-specific assumptions that are difficult to replicate or systematically compare. As a result, claims of privacy or anonymity robustness may not generalize beyond the original evaluation context.

In this poster, we propose a structured framework for the reproducible evaluation of privacy and anonymity guarantees. The framework decomposes privacy evaluation into a set of explicit dimensions, including adversary knowledge, interaction model, auxiliary information, adaptivity, and observability of leakage. We argue that variations along these dimensions materially affect evaluation outcomes, and we demonstrate how the absence of standardized reporting hinders meaningful comparison across systems and studies.

Our goal is not to introduce a new privacy mechanism, but to provide a conceptual foundation and reporting schema that improves rigor, transparency, and reproducibility in privacy and anonymity research. We position this work as a discussion-oriented contribution aimed at fostering more systematic and comparable privacy evaluations in future security research.

Keywords

Privacy; Anonymity; Reproducibility; Evaluation Methodology; Threat Modeling; Security Metrics

1 Introduction

Privacy and anonymity have evolved from niche properties into foundational requirements for modern distributed systems. From mix-networks and anonymous communication systems (ACS) to privacy-preserving machine learning, the security community continues to produce novel mechanisms designed to protect user identity and data confidentiality. However, while the design of these mechanisms has matured, the empirical evaluation of their guarantees remains inconsistent and difficult to reproduce [5, 7].

Unlike cryptographic primitives, which often benefit from binary failure criteria (e.g., a cipher is computationally secure or it is broken), system-level privacy is typically evaluated along a continuum of leakage. Researchers often demonstrate robustness by simulating attacks under specific network conditions or adversarial capabilities. Unfortunately, these evaluations are frequently grounded in implicit threat models [11]. For instance, a system proven secure against a local passive adversary (who observes a fraction of network traffic) may catastrophically fail against a global active adversary (who manipulates traffic timing), yet papers often conflate these scenarios or fail to explicitly scope their claims [5].

This lack of standardization creates a "comparability crisis." Without a unified schema for reporting adversarial assumptions, such as the adversary's knowledge of the network topology, their ability to inject auxiliary traffic, or their adaptability, it is essentially impossible to compare the privacy guarantees of System A against System B [5, 9]. As a result, experimental artifacts are often irreproducible, and "robustness" becomes a property of the experimental setup rather than the system itself [7].

In this work, we propose a structured framework for the reproducible evaluation of privacy and anonymity. We decompose privacy assessment into five explicit dimensions: **(1) Adversary Knowledge**, **(2) Interaction Model**, **(3) Auxiliary Information**, **(4) Adaptivity**, and **(5) Observability**. Our contribution is not a new privacy protocol, but rather a systematization of evaluation methodology. We argue that by explicitly defining these dimensions, the community can move toward a standardized reporting schema that fosters transparency, enables rigorous peer review, and allows for meaningful cross-study comparison.

2 The Comparability Gap

The core challenge in empirical privacy research is not the absence of mathematical metrics - definitions ranging from k-anonymity to differential privacy are well-established, but rather the inconsistent application of the adversarial context in which these metrics are measured. This inconsistency manifests primarily through *implicit threat models* and *parameter divergence*, rendering cross-study comparisons effectively impossible [5].

First, the reliance on **implicit threat models** leads to incommensurate comparisons between systems. For example, in the evaluation of anonymous communication networks, one study may assume a *static adversary* who observes a fixed 20% of network nodes [3], while a competing study assumes a *dynamic adversary* capable of adaptively compromising nodes based on traffic volume [6]. A system that appears robust under the first model may fail catastrophically under the second. However, without a standardized reporting schema, these critical distinctions are often buried in supplementary materials or omitted entirely, allowing weaker systems to claim equivalent "robustness" to stronger ones [11].

Second, **parameter divergence** in experimental setups creates a barrier to reproducibility. Recent surveys in privacy-preserving machine learning (PPML) indicate that attack success rates are highly sensitive to auxiliary information, such as the specific distribution of the training data [7]. When a study claims a high defense rate against membership inference attacks but fails to

,
2018.

specify the exact prior knowledge granted to the adversary [10], the result becomes an artifact of that specific experiment rather than a generalizable property of the system. This leads to a fragmented literature where "state-of-the-art" is a moving, undefined target, and contradictory claims coexist without resolution.

3 The Proposed Evaluation Framework

To resolve these disparities, we introduce a structured evaluation framework that decomposes privacy assessments into five explicit dimensions. Our goal is to shift the community from ad-hoc, implicit threat modeling to a standardized reporting schema. By plotting an evaluation along these axes, researchers can create a "fingerprint" of their security claims, ensuring that omitted variables (such as adversary adaptivity) are explicitly acknowledged rather than silently ignored.

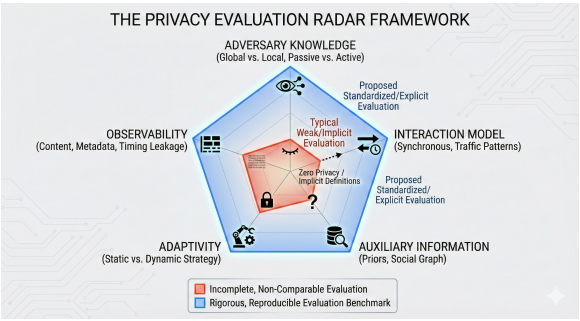

Figure 1: The Privacy Evaluation Radar Framework contrasting typical weak evaluations (Red) against the proposed explicit standard (Blue).

3.1 Dimension A: Adversary Knowledge

This dimension captures *what* the adversary knows about the system topology and parameters. It distinguishes between:

- **Zero-Knowledge:** The adversary knows only the public protocol specification but lacks insight into the current network state.
- **Partial/Local Knowledge:** The adversary observes a subset of the network (e.g., 20% of mix-nodes) or possesses statistical distributions of user inputs [3].
- **Full/Global Knowledge:** The adversary has a complete "God-view" of the infrastructure and algorithms.

Reporting Requirement: Evaluations must explicitly state whether robustness holds against a global adversary or is contingent on the adversary's limited view.

3.2 Dimension B: Interaction Model

This dimension defines *how* the adversary interacts with the target system.

- **Passive:** The adversary is restricted to eavesdropping on network traffic or querying public databases without altering state.
- **Active:** The adversary can inject, drop, replay, or modify messages to induce errors or timing delays (e.g., tagging attacks in Tor) [1].
- **Malicious/Byzantine:** The adversary controls internal system nodes and can deviate from the protocol arbitrarily.

Reporting Requirement: Claims of "anonymity" must specify if they hold against active traffic shaping or only passive observation.

3.3 Dimension C: Auxiliary Information

This dimension quantifies the external knowledge the adversary leverages to de-anonymize users. A rigorous evaluation must explicitly define the "priors" available to the attacker.

- **Distributional Priors:** Knowledge of the user's likely behavior (e.g., "User A usually logs in at 9 AM").
- **Social Graph Knowledge:** Information regarding who communicates with whom, which can be used to break anonymity even in encrypted channels [2].

Reporting Requirement: Evaluations must state whether the system protects users with unique, identifiable habits (outliers) or only "average" users with uniform behavior.

3.4 Dimension D: Adaptivity

Most evaluations assume a static adversary. This dimension measures the adversary's ability to react to the defense *in real-time*.

- **Static Strategy:** The attacker decides on a strategy (e.g., "monitor Node 5") before the experiment begins and sticks to it regardless of system state.
- **Adaptive Strategy:** The attacker updates their strategy based on observed leakage (e.g., "I saw traffic on Node 5, so I will now compromise Node 6") [8].

Reporting Requirement: Security claims must clarify if they hold against an adaptive adversary who optimizes their attack mid-stream.

3.5 Dimension E: Observability of Leakage

Finally, we must standardize *what* constitutes a failure.

- **Content Leakage:** The adversary recovers the plaintext message.
- **Metadata Leakage:** The adversary learns the sender, receiver, message size, or route length.
- **Timing Leakage:** The adversary correlates entry and exit events based on latency.

Reporting Requirement: A system is rarely "secure" in binary terms; evaluations must specify the *type* of leakage tolerated (e.g., "We hide content but leak timing").

3.6 Evaluation Workflow

To apply this framework, researchers should perform the following tasks:

(1) **Baseline Selection:** Identify the target privacy metric (e.g., k-anonymity).
(2) **Dimension Mapping:** Assign a value (Low/Medium/High) to each of the five dimensions based on the experimental setup.
(3) **Sensitivity Analysis:** Vary one dimension (e.g., Adaptivity) while holding others constant to identify "collapse points" in the defense.
(4) **Comparative Reporting:** Plot the resulting "security boundary" to enable cross-study comparison.

4 Case Study: Website Fingerprinting on Tor

To demonstrate the utility of this framework, we examine a well-documented crisis in privacy evaluation: **Website Fingerprinting (WF) attacks on Tor**. WF attacks attempt to identify encrypted web traffic by analyzing packet sizes and timing. Over the last decade, evaluations of these attacks have produced wildly contradictory claims regarding Tor's robustness.

We apply our framework to compare two classes of existing studies:

Evaluation A: The "Closed World" Assumption
Many early ML-based studies evaluated attacks assuming the user visits only a small, fixed set of monitored websites (e.g., top 100 Alexa sites).

- **Dimension C (Auxiliary Info):** Implicitly set to "Closed World." The adversary knows the user *must* be visiting one of the 100 known sites.
- **Reported Result:** Attacks claimed > 90% accuracy, leading to the conclusion that Tor is fundamentally broken against traffic analysis.

Evaluation B: The "Open World" Reality
Later critiques re-evaluated these attacks in a realistic setting where users visit thousands of sites [4].

- **Dimension C (Auxiliary Info):** Explicitly set to "Open World." The adversary must distinguish monitored sites from the vast "background noise" of the web.
- **Reported Result:** Due to the base rate fallacy, the False Positive Rate (FPR) exploded. Practical accuracy dropped to < 3%, leading to the conclusion that Tor is largely robust against these specific attacks [4].

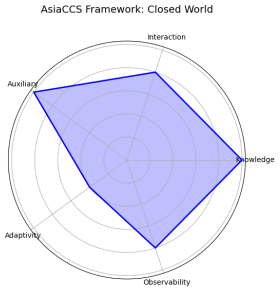

Figure 2: Closed World: High precision artifacts appear as a stable area.

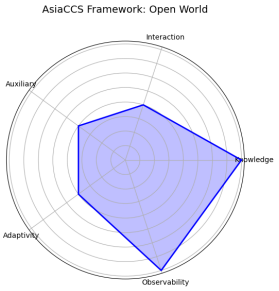

Figure 3: Open World: Real-world noise causes "Security Collapse."

The Framework's Impact
Without a standardized schema, Evaluation A was able to claim "state-of-the-art" results that were essentially artifacts of the experimental design. Under our proposed framework, Evaluation A would be required to explicitly report **Dimension C** as *"High Priors / Closed World."* This transparency would immediately signal to reviewers and practitioners that the reported "90% accuracy" does not generalize.

Reproducibility: To support the peer review process, the complete evaluation framework, including the comparative dataset and radar chart scripts, is available at: https://github.com/security-researcher-101/asiaccs-2026

5 Conclusion

As privacy-enhancing technologies mature, the methods used to evaluate them must evolve from ad-hoc experiments to rigorous science. We argue that the current "comparability crisis" in privacy research is driven largely by implicit threat models and underspecified adversary capabilities. This lack of transparency allows weak evaluations to masquerade as robust proofs of security, hindering progress and confusing practitioners [5].

In this work, we proposed a structured evaluation framework that decomposes privacy assessments into five explicit dimensions: **Adversary Knowledge, Interaction Model, Auxiliary Information, Adaptivity,** and **Observability**. We demonstrated, through the case of Website Fingerprinting on Tor, how variations along a single dimension (Auxiliary Information) can invert the conclusions of an entire sub-field [4].

We position this framework not as a rigid checklist, but as a "reporting schema"—a standardized language for describing the security boundary of a system. By adopting this schema, the security community can ensure that future contributions are reproducible, transparent, and meaningfully comparable. We invite the AsiaCCS community to discuss, refine, and adopt these dimensions to establish a new standard for empirical rigor in privacy research.

References

[1] George Danezis. 2004. The Traffic Analysis of Continuous-Time Mixes. In *Proceedings of Privacy Enhancing Technologies (PET)*. Springer.
[2] George Danezis and Prateek Mittal. 2009. SybilInfer: Detecting Sybil Nodes using Social Networks. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*. The Internet Society.
[3] Roger Dingledine, Nick Mathewson, and Paul Syverson. 2004. Tor: The Second-Generation Onion Router. In *Proceedings of the 13th USENIX Security Symposium*. USENIX Association.
[4] Marc Juarez, Sadia Afroz, Gunes Acar, Claudia Diaz, and Rachel Greenstadt. 2014. A Critical Evaluation of Website Fingerprinting Attacks. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM.
[5] Christiane Kuhn, Martin Beck, Stefan Schiffner, Eduard Jorswieck, and Thorsten Strufe. 2019. On Privacy Notions in Anonymous Communication. *Proceedings on Privacy Enhancing Technologies (PoPETs)* 2019, 2 (2019), 105–125.
[6] Steven J. Murdoch and George Danezis. 2005. Low-Cost Traffic Analysis of Tor. In *2005 IEEE Symposium on Security and Privacy (SP)*. IEEE.
[7] Daniel Olszewski, Christopher A. Choquette-Choo, Lukas Ifflaender, Florian Tramèr, Nicholas Carlini, and Carmela Troncoso. 2023. A Reproducibility Study of Machine Learning Papers in Tier 1 Security Conferences. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM.
[8] Simon Oya, Carmela Troncoso, and Fernando Pérez-González. 2017. Back to the Drawing Board: Revisiting the Design of Optimal Location Privacy-Preserving Mechanisms. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM. (Reference for Adaptive/Strategic Adversaries).
[9] Andreas Pfitzmann and Marit Hansen. 2010. *A terminology for talking about privacy by data minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management.* Technical Report v0.34. Technische Universität Dresden. https://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf.
[10] Reza Shokri, George Theodorakopoulos, Jean-Yves Le Boudec, and Jean-Pierre Hubaux. 2011. Quantifying Location Privacy. In *2011 IEEE Symposium on Security and Privacy (SP)*. IEEE.
[11] Nik Unger, Sergej Dechand, Joseph Bonneau, Sascha Fahl, Henning Perl, Ian Goldberg, and Matthew Smith. 2015. SoK: Secure Messaging. In *Proceedings of the 2015 IEEE Symposium on Security and Privacy (SP)*. IEEE, 232–249.