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

Anonymous Author(s)

Abstract

Privacy and anonymity are fundamental goals of modern systems, yet their evaluation is often inconsistent and difficult to reproduce. Prior work frequently relies on implicit threat models and loosely defined adversary assumptions, making results hard to compare across studies. As a result, reported privacy guarantees often hold only within narrow, system-specific settings.

We propose a structured framework for the reproducible evaluation of privacy guarantees. The framework explicitly separates key evaluation dimensions, including adversary knowledge, interaction models, auxiliary information, and adaptivity. We show that variations along these dimensions can substantially affect evaluation outcomes, and that the lack of standardized reporting prevents meaningful comparison between systems.

Rather than introducing a new defense mechanism, this work provides a conceptual foundation and reporting schema to improve rigor and transparency in privacy evaluations. Our goal is to support more systematic and comparable privacy research by grounding claims in clearly stated, reproducible assumptions.

Keywords

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

1 Introduction

Privacy and anonymity have transitioned from niche features to foundational requirements for modern distributed systems. While the design of privacy-preserving mechanisms, ranging from mix-networks to secure machine learning, has matured significantly, the empirical evaluation of their guarantees remains inconsistent and notoriously difficult to reproduce [5, 7].

Unlike cryptographic primitives, which generally adhere to binary security definitions where a cipher is either secure or broken, system-level privacy operates on a continuum of leakage. Researchers typically demonstrate robustness by simulating attacks under specific conditions, yet these evaluations often rest on implicit threat models [11]. A system secure against a local passive observer may fail catastrophically against a global active adversary. Conflating these distinct threat landscapes, or failing to explicitly scope claims, undermines the validity of the findings [5].

This ambiguity precipitates a "comparability crisis." Absent a unified reporting schema that defines critical variables such as topological knowledge, traffic injection capabilities, or adversarial adaptivity, it is effectively impossible to benchmark System A against System B [5, 9]. Consequently, experimental artifacts often become irreproducible, and "robustness" reflects the specific constraints of an experimental setup rather than the intrinsic resilience of the system itself [7].

To address this challenge, we propose a structured framework for the reproducible evaluation of privacy guarantees. We decompose 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 protocol, but a systematization of methodology. By codifying these parameters, we aim to establish a standard that promotes transparency, ensures rigorous peer review, and enables meaningful cross-study comparison.

2 The Comparability Gap

The primary obstacle in empirical privacy research is not a lack of mathematical rigor. Metrics such as k-anonymity and differential privacy are well established. Instead, the difficulty lies in the inconsistent application of the adversarial context in which these metrics are measured. This inconsistency surfaces primarily through *implicit threat models* and *parameter divergence*, which renders cross-study comparisons effectively impossible [5].

Reliance on **implicit threat models** produces incommensurate comparisons between systems. Consider the evaluation of anonymous communication networks. One study might assume a *static adversary* who observes a fixed percentage of network nodes [3]. A competing study might assume 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. Yet, absent a standardized reporting schema, these critical distinctions are often buried in supplementary materials or omitted entirely. This ambiguity allows weaker systems to claim equivalent "robustness" to stronger ones [11].

Furthermore, **parameter divergence** in experimental setups erects 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 training data [7]. If a study claims a high defense rate against membership inference attacks but fails to specify the exact prior knowledge granted to the adversary [10], the result remains an artifact of that specific experiment rather than a generalizable property of the system. This fragmentation reduces "state-of-the-art" to a moving target where contradictory claims coexist without resolution.

3 The Proposed Evaluation Framework

To address these disparities, we introduce a structured framework that decomposes privacy assessments into five explicit dimensions. Our objective is to shift the community from ad hoc, implicit threat modeling toward a standardized reporting schema. By plotting an

,
2018.

evaluation along these axes, researchers create a "fingerprint" of their security claims. This ensures that omitted variables, such as adversary adaptivity, are explicitly acknowledged rather than silently ignored.

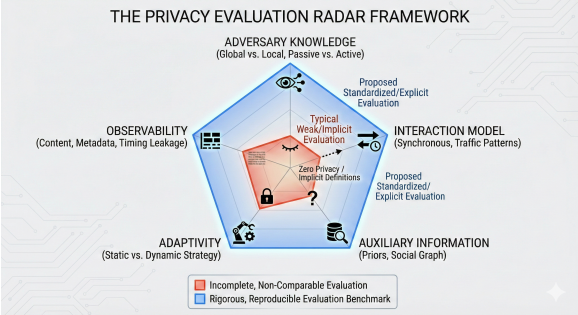

Figure 1: The Privacy Evaluation Radar Framework contrasting typical weak evaluations (Red) against the proposed explicit standard (Blue).

3.1 Dimension A: Adversary Knowledge

This dimension captures the extent of the adversary's insight into 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, such as 20% of mix-nodes, or possesses statistical distributions of user inputs [3].
- **Full/Global Knowledge:** The adversary possesses a complete, omniscient 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 the adversary's capacity to interact 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, typified by 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 an adversary leverages to de-anonymize users. Rigorous evaluation demands an explicit definition of the "priors" available to the attacker.

- **Distributional Priors:** Knowledge of the user's likely behavior, such as specific login times.
- **Social Graph Knowledge:** Information regarding communication patterns, 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 default to 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, like monitoring Node 5, before the experiment begins and adheres to it regardless of system state.
- **Adaptive Strategy:** The attacker updates their strategy based on observed leakage, such as compromising a new node after detecting traffic elsewhere [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 the definition of 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: Systems are rarely "secure" in binary terms. Evaluations must specify the *type* of leakage tolerated, for example, hiding content while leaking timing.

3.6 Evaluation Workflow

To apply this framework, researchers engage in the following tasks:

(1) **Baseline Selection:** Identify the target privacy metric, such as k-anonymity.
(2) **Dimension Mapping:** Assign a value (Low, Medium, or High) to each of the five dimensions based on the experimental setup.
(3) **Sensitivity Analysis:** Vary one dimension, such as 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 practical utility of this framework, we examine a well-documented disparity in privacy evaluation: **Website Fingerprinting (WF) attacks on Tor**. WF attacks attempt to identify encrypted web traffic through the analysis of packet sizes and timing sequences. Over the last decade, evaluations of these attacks have yielded contradictory claims regarding the robustness of Tor.

We apply our framework to contrast two distinct classes of existing studies:

Evaluation A: The "Closed World" Assumption
Early machine learning-based studies frequently evaluated attacks under the assumption that the user visits only a small, fixed set of monitored websites, such as the Alexa Top 100.

- **Dimension C (Auxiliary Info):** Implicitly constrained to a "Closed World." The adversary operates with the knowledge that the user *must* be visiting one of the 100 known sites.
- **Reported Result:** These attacks achieved accuracy exceeding 90%. Consequently, authors concluded that Tor was fundamentally vulnerable to traffic analysis.

Evaluation B: The "Open World" Reality
Subsequent critiques reassessed these attacks in a realistic setting where users visit thousands of diverse sites [4].

- **Dimension C (Auxiliary Info):** Explicitly expanded to an "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) increased precipitously. Operational precision fell to less than 3%, which supports the conclusion that Tor remains largely robust against these specific attacks [4].

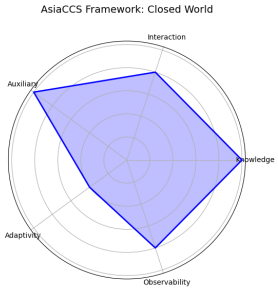

Figure 2: Closed World: High precision artifacts appear as a stable area.

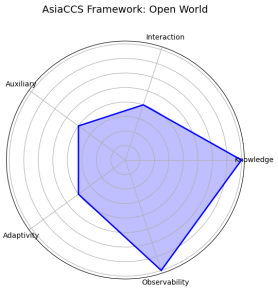

Figure 3: Open World: Real-world noise causes "Security Collapse."

The Framework's Impact
Absent a standardized schema, Evaluation A claimed "state-of-the-art" results that were, in reality, artifacts of the experimental design. Under our proposed framework, Evaluation A would be mandated to explicitly report **Dimension C** as *"High Priors / Closed World."* Such transparency serves as an immediate signal to reviewers and practitioners that the reported 90% accuracy does not generalize to real-world deployment.

Reproducibility: To facilitate 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 methodologies used to evaluate them must transition from ad hoc experimentation to rigorous science. We contend that the current "comparability crisis" in privacy research stems largely from implicit threat models and underspecified adversary capabilities. This opacity permits weak evaluations to pass as robust proofs of security, obstructing progress and confusing practitioners [5].

To remedy this, we propose a structured framework that decomposes privacy assessments into five explicit dimensions: **Adversary Knowledge, Interaction Model, Auxiliary Information, Adaptivity,** and **Observability**. Using the case of Website Fingerprinting on Tor, we demonstrated how variations along a single dimension, specifically Auxiliary Information, can invert the conclusions of an entire subfield [4].

We offer this framework not as a rigid checklist, but as a "reporting schema," a standardized language for describing the security boundary of a system. Adopting this schema ensures that future contributions remain 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.

Generative AI tools (Google Gemini and OpenAI ChatGPT) were used to assist in the drafting and linguistic refinement of this text.

References

[1] George Danezis. 2004. The Traffic Analysis of Continuous-Time Mixes. In *Proceedings of Privacy Enhancing Technologies (PET)*. Springer. https://www.freehaven.net/anonbib/cache/danezis:pet2004.pdf

[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. http://www0.cs.ucl.ac.uk/staff/g.danezis/papers/sybilinfer.pdf

[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. https://svn.torproject.org/svn/projects/design-paper/tor-design.pdf

[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. https://nymity.ch/tor-dns/pdf/Juarez2014a.pdf

[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. https://petsymposium.org/popets/2019/popets-2019-0022.pdf

[6] Steven J. Murdoch and George Danezis. 2005. Low-Cost Traffic Analysis of Tor. In *2005 IEEE Symposium on Security and Privacy (SP)*. IEEE. https://murdoch.is/papers/oakland05torta.pdf

[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. https://www.cise.ufl.edu/~butler/pubs/ccs23-olszewski.pdf

[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. http://www.carmelatroncoso.com/papers/Oya-CCS17.pdf (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. https://www.ieee-security.org/TC/SP2011/PAPERS/2011/paper016.pdf

[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. https://www.ieee-security.org/TC/SP2015/papers-archived/6949a232.pdf