
Practical Bayes-Optimal Membership Inference Attacks

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 We develop practical and theoretically grounded membership inference attacks
2 (MIAs) against both independent and identically distributed (i.i.d.) data and graph-
3 structured data. Building on the Bayesian decision-theoretic framework of [1],
4 we derive the Bayes-optimal membership inference rule for node-level MIAs
5 against graph neural networks, addressing key open questions about optimal query
6 strategies in the graph setting. We introduce BASE and G-BASE, computationally
7 efficient approximations of the Bayes-optimal attack. G-BASE achieves superior
8 performance compared to previously proposed classifier-based node-level MIA
9 attacks. BASE, which is also applicable to non-graph data, matches or exceeds
10 the performance of prior state-of-the-art MIAs, such as LIRA and RMIA, at a
11 significantly lower computational cost. Finally, we show that BASE and RMIA
12 are equivalent under a specific hyperparameter setting, providing a principled,
13 Bayes-optimal justification for the RMIA attack.

14 1 Introduction

15 Machine learning models are known to leak information about their training data [2–5], driving
16 growing interest in understanding and quantifying such leakage. Due to the complexity of modern
17 models, formally characterizing their information leakage remains a formidable challenge. As a
18 result, privacy is often assessed empirically via so-called *privacy auditing*—designing attacks to
19 identify data leakage and guide mitigation strategies.

20 Membership inference attacks (MIAs), where an adversary seeks to infer whether a specific data
21 point was included in the training set of a model, represent the most fundamental privacy attack.
22 Importantly, MIAs already pose a serious privacy threat: disclosing the mere presence of a data
23 point in the training set may constitute a serious privacy violation and directly contravenes privacy
24 regulations such as the GDPR. The European Data Protection Board explicitly cites MIAs in its
25 guidance on when a machine learning model can be considered anonymous [6]. Moreover, MIAs can
26 serve as a key component in data reconstruction attacks by filtering out unlikely candidates, thereby
27 narrowing the search space [7], and can be used to establish lower bounds on the privacy guarantees
28 of differentially-private algorithms [8].

29 State-of-the-art MIAs often rely on *shadow models*—auxiliary models trained under similar condi-
30 tions as the target model—to empirically characterize behavioral differences between training and
31 non-training data. Existing attacks fall into two main categories: classifier-based and statistic-based.
32 Classifier-based attacks use features from shadow models, such as losses or logits, to train a binary
33 classifier [2, 9, 10]. Statistic-based attacks instead compute statistical metrics using signals from
34 both the target and shadow models [1, 11–13]. A Bayes-optimal strategy to membership inference
35 was proposed in [1], but it is intractable in practice, necessitating approximations. State-of-the-art
36 methods such as LIRA [11] and RMIA [13], instead, are based on a hypothesis testing formu-

lation [14]. While these approaches achieve strong empirical performance and often outperform classifier-based attacks, their theoretical connection to the Bayes-optimal rule remains unclear.

Most existing work on MIAs assumes data points are independent and identically distributed (i.i.d.). Recently, however, MIAs have been studied in the context of graph-structured data, particularly against graph neural networks (GNNs). In this setting, message passing introduces structural dependencies—each node or edge can influence many others—making membership signals harder to isolate and challenging key assumptions underlying classical attacks. MIAs on graph-structured data can be grouped by the information they aim to recover: (i) node-level attacks attempt to infer whether a specific node was part of the training set [15–19]; (ii) edge-level attacks target the presence of specific edges [20–24]; and (iii) graph-level attacks aim to determine whether an entire graph instance was used in the training [25–27]. At the node level, existing approaches are exclusively classifier-based. Given the success of statistic-base methods on the i.i.d. data, extending such strategies to graph data may lead to stronger attacks. However, this remains an open problem.

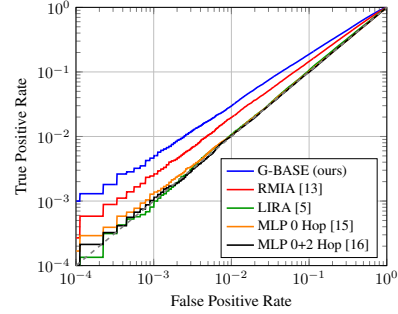


Figure 1: Online attack ROC on Flickr dataset averaged over 10 GCN target models.

Our contribution. Motivated by the need for practical, theoretically grounded privacy auditing tools, we develop novel MIAs for both i.i.d. and graph data. Specifically, we design Bayes-optimal attacks building on the Bayesian decision-theoretic framework of [1], yielding attacks that are both effective and computationally efficient. The proposed attacks match or outperform the state-of-the-art while significantly reducing computational overhead. Our key contributions include:

- We derive the Bayes-optimal decision rule for node-level membership inference on graph data, extending the results of Sablayrolles et al. [1] beyond the i.i.d. setting. This result formalizes how graph structure should be exploited in membership inference against GNNs.
- Guided by this Bayes-optimal rule, we propose G-BASE, a practical attack on GNNs that achieves state-of-the-art performance on graph data, significantly outperforming existing classifier-based methods.
- We show that the Bayes-optimal MIA for i.i.d. data in [1] is a special case of our Bayes-optimal decision rule for graph data.
- For i.i.d. data, we propose BASE, a practical Bayes-optimal attack achieving state-of-the-art performance while being significantly more cost-efficient than prior methods such as RMIA [13] and LIRA [11]. In particular, BASE outperforms RMIA in the offline setting while performing on par in the online setting.
- For i.i.d. data, we reveal a close connection between BASE and RMIA in the online setting: the two are equivalent (for a specific value of RMIA’s threshold γ) up to a monotonically increasing transformation. However, BASE achieves the same performance with significantly lower computational complexity, requiring much fewer model queries.

2 Related Work

MIAs for i.i.d. data. Relevant works include [2], [1], [14], [11] (LIRA), and [13] (RMIA). Shokri et al. [2] introduced shadow training for black-box MIAs: the adversary trains multiple shadow models on known datasets to mimic the target model, then trains an attack model on their labeled inputs and outputs. Sablayrolles et al. [1] proposed a Bayes-optimal framework in which membership inference amounts to computing the posterior probability that a data point is in the training set, given the trained model. Under mild assumptions, the optimal inference depends only on the loss, motivating black-box attacks where the attacker can observe only the model’s output (e.g., the loss) on the target sample. Ye et al. [14] cast MIAs as a hypothesis testing game: a challenger samples a data point from either the training set or the rest of the data population (with equal probability), and an adversary, with access to the data population, must decide between two hypothesis—whether the model was trained on a dataset including or excluding the target point. State-of-the-art black-box attacks, such as LIRA [11] and RMIA [13], build on this formulation. LIRA performs a likelihood-ratio test between the two hypotheses, representing them via the distributions of losses (or logits) on the target data point. LIRA assumes the (transformed) logits exhibit Gaussian-like behavior, resulting

in a parametric test. In RMIA, the test compares the case where the target data point is in the training set to the case where it is replaced by an auxiliary sample drawn randomly from the population. The test score is the tail probability of the resulting likelihood ratio.

MIAs for graph data. Existing node-level attacks [15–19] all adopt a classifier-based approach. These attacks involve three phases: training shadow models, training a binary-classifier attack model, and performing membership inference. The attack model takes as input features the vector derived from the shadow model’s outputs—either posterior probabilities [15–17, 19] or predicted labels [18]—for nodes that were included or not in the training set. Despite their promise, statistic-based methods have not yet been explored for node-level MIAs.

3 Preliminaries

Graph notation. For graph data, we consider a graph denoted by $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X}, \mathbf{Y})$, where \mathcal{V} is the set of n nodes, \mathcal{E} the set of edges, $\mathbf{X} \in \mathbb{R}^{n \times d}$ the node feature matrix, and $\mathbf{Y} \in \mathbb{R}^{n \times c}$ the one-hot encoded node-label matrix, such that each node $v \in \mathcal{V}$ has an associated feature-label pair $(\mathbf{x}_v, \mathbf{y}_v)$, which are assumed to be sensitive data. We denote by $\mathbf{A} \in \mathbb{R}^{n \times n}$ the adjacency matrix of graph \mathcal{G} , where $A_{uv} = 1$ if $(u, v) \in \mathcal{E}$ and 0 otherwise, and by $\mathcal{N}(v) = \{u : (u, v) \in \mathcal{E}, u \neq v\}$ the set of neighbors of node v . For convenience, in the analysis sections, we will refer to a graph \mathcal{G} as $\mathcal{G} = (\mathbf{X}, \mathbf{Y}, \mathbf{A})$, where the sets of nodes and edges are implicitly defined by \mathbf{X} and \mathbf{A} , respectively.

Graph neural networks (GNNs). GNNs [28–31] learn node embeddings that capture both structural and feature information. These embeddings are then used for various downstream tasks such as node classification. GNNs operate via message passing, where nodes aggregate information from their neighbors to update their embeddings. Formally, the embedding of node $v \in \mathcal{V}$ at level $\ell + 1$ of an L -layer GNN is computed as

$$\mathbf{h}_v^{(\ell+1)} = \text{UPDATE}\left(\mathbf{h}_v^{(\ell)}, \text{AGGREGATE}\left(\left\{\mathbf{h}_u^{(\ell)}, u \in \mathcal{N}(v)\right\}\right)\right), \quad (1)$$

where both AGGREGATE, a permutation-invariant function, and UPDATE are differentiable and $\mathbf{h}_v^{(0)}$ is the input feature vector of node v . The computation of $\mathbf{h}_v^{(L)}$, the final embedding of node v , thus depends on its L -hop neighborhood, which we denote by $\mathcal{N}_L(v)$. For notational convenience, we denote the final embedding of node v by $\mathbf{z}_v = \mathbf{h}_v^{(L)}$ and refer to it simply as the node embedding.

For node classification, a softmax operation is typically applied to the node embeddings to produce class probabilities,

$$P(\mathbf{y}_v | \boldsymbol{\theta}, \mathbf{X}, \mathbf{A}) \approx \text{softmax}(\mathbf{z}_v)_y \equiv f_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{A})_{vy} \quad (2)$$

where $f_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{A})_v$ denotes the softmax-normalized prediction vector for node v , computed by a GNN $f_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta}$, and y denotes the index of the nonzero element of \mathbf{y}_v . The model is typically trained by minimizing the negative log-likelihood loss $\ell(f_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{A})_v, \mathbf{y}_v) = -\log P(\mathbf{y}_v | \boldsymbol{\theta}, \mathbf{X}, \mathbf{A})$ (i.e., the cross-entropy loss for node classification) over the labeled nodes, i.e., minimizing

$$\mathcal{L}(\boldsymbol{\theta}, \mathbf{X}, \mathbf{Y}, \mathbf{A}) = \sum_{v \in \mathcal{V}} \ell(f_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{A})_v, \mathbf{y}_v) = - \sum_{v \in \mathcal{V}} \log P(\mathbf{y}_v | \boldsymbol{\theta}, \mathbf{X}, \mathbf{A}). \quad (3)$$

4 Bayes-Optimal Node-Level MIAs Against GNNs

We extend the Bayesian membership inference strategy in [1] to GNNs. We begin by framing the MIA as an indistinguishability game and then proceed to derive the Bayes-optimal decision rule for membership inference. Finally, we discuss tractable approximations and sampling strategies that yield powerful and practical MIAs.

4.1 Node Membership Inference Attack Game

Following [11, 14], we formulate the MIA as a game between a challenger and an adversary.

Definition 1. (Membership inference game on graph data)

1. The challenger samples a subset of graph \mathcal{G} , $\mathcal{G}_{\text{train}} = (\mathcal{V}_{\text{train}}, \mathcal{E}_{\text{train}}) \subset \mathcal{G}$, where $\mathcal{V}_{\text{train}} \in \mathcal{V}$ and $\mathcal{E}_{\text{train}} = \{(u, v) : u, v \in \mathcal{V}_{\text{train}}, (u, v) \in \mathcal{E}\}$, and trains a GNN model $f_{\boldsymbol{\theta}}$ on $\mathcal{G}_{\text{train}}$. Let m_v denote

134 the membership status of node $v \in \mathcal{V}$,

$$m_v = \begin{cases} 1, & v \in \mathcal{V}_{\text{train}} \\ 0, & v \notin \mathcal{V}_{\text{train}} \end{cases}. \quad (4)$$

- 135 2. The challenger flips a fair coin to generate a bit b . If $b = 0$, it samples a target node v node
136 from $\mathcal{G} \setminus \mathcal{G}_{\text{train}}$. Otherwise, it samples a target node v from $\mathcal{G}_{\text{train}}$.
- 137 3. The challenger gives the adversary the full graph $\mathcal{G} = (\mathbf{X}, \mathbf{Y}, \mathbf{A})$, the target node v , and
138 black-box access to the trained model f_θ .
- 139 4. The adversary may also have additional side information \mathcal{H} (e.g., knowledge about the training
140 algorithm or model architecture). Using this information, the adversary performs a MIA $\hat{m}_v \leftarrow$
141 $\text{MIA}(f_\theta, v, \mathcal{G}, \mathcal{H})$, where \hat{m}_v is an estimate of the membership status of node v , m_v .
- 142 5. The attack is successful on a target node v if $\hat{m}_v = m_v$.

143 **Threat model.** We assume that the target graph $\mathcal{G}_{\text{train}} \subset \mathcal{G}$ is sampled from a larger, fixed graph \mathcal{G}
144 that is accessible to the adversary. This setup offers a flexible and practical way to model the data
145 distribution. Notably, this corresponds to the *train on subgraph, test on full* setting introduced in [15].
146 When \mathcal{G} is large relative to $\mathcal{G}_{\text{train}}$, this setup corresponds to the *data population pool* assumption widely
147 adopted in membership inference [14, 13, 11]. While giving the adversary access to \mathcal{G} simplifies the
148 analysis of the Bayes-optimal attack, it also results in a stronger adversary. Nevertheless, this worst
149 case assumption is suitable for privacy auditing, where conservative evaluations are preferred. We
150 adopt a black-box setting, where the adversary has unlimited query access to the target model and can
151 generate soft predictions for any valid input graph. Knowledge of the model’s architecture, training
152 algorithm, and objective function is encompassed in \mathcal{H} . In contrast, a white-box setting would grant
153 the adversary direct access to model parameters or activations.

154 GNN training can be categorized into *transductive* and *inductive*. In the transductive setting, only
155 part of the graph is labeled and used for loss computation. However, message-passing is performed
156 over all nodes, producing an embedding for each node. The goal of this semi-supervised learning
157 approach is typically to predict the labels of the unlabeled nodes. The inductive setting, on the other
158 hand, corresponds to fully supervised learning. In this setting, message-passing is only performed
159 between labeled nodes. The goal is then to generalize to unseen nodes. As noted in [17], the inductive
160 setting is the most relevant, as it provides a clear distinction between member and non-member
161 nodes. In contrast, the transductive setting—where the model is primarily used by the data holder to
162 predict labels for unlabeled nodes and remains under their control—does not raise significant privacy
163 concerns. We therefore focus our analysis and experiments on the inductive setting.

164 4.2 The Bayes-Optimal Decision Rule

165 Assume that the target node whose membership status we aim to infer is node $v \in \mathcal{I}$, and let $\mathcal{N}_L(v)$
166 be its L -hop neighborhood. Also, let $\mathcal{M} = \{m_u : u \in \mathcal{V}\}$ denote the membership indicator variables
167 of all n nodes in the graph \mathcal{G} , and $\tilde{\mathcal{M}} = \mathcal{M} \setminus \{m_v\}$ represent the membership statuses of all nodes
168 except the target node v . Given the attack game and threat model defined in Section 4.1, a Bayesian
169 adversary seeks to compute the posterior probability that the target node is in the training set, i.e.,
170 $P(m_v = 1 | \theta, \mathcal{G})$. The following theorem provides a closed-form expression for this posterior.

171 **Theorem 1.** Given the graph $\mathcal{G} = (\mathbf{X}, \mathbf{Y}, \mathbf{A})$ and the L -layer GNN model θ , the posterior probability
172 $P(m_v = 1 | \theta, \mathcal{G})$ is given by

$$P(m_v = 1 | \theta, \mathcal{G}) = \mathbb{E}_{\tilde{\mathcal{M}} \sim P(\tilde{\mathcal{M}} | \theta, \mathcal{G})} \left[\sigma \left(-S_{\mathcal{L}}(f_\theta, v, \tilde{\mathcal{M}}, \mathcal{G}) - \log \int e^{-S_{\mathcal{L}}(f_\phi, v, \tilde{\mathcal{M}}, \mathcal{G})} p(\phi | \tilde{\mathcal{M}}, \mathcal{G}) d\phi + \log \frac{\lambda}{1 - \lambda} \right) \right], \quad (5)$$

173 where

$$S_{\mathcal{L}}(f_\theta, v, \tilde{\mathcal{M}}, \mathcal{G}) = \ell(f_\theta(\mathbf{X}, \mathbf{A}_{\mathcal{M}})_v, \mathbf{y}_v) + \Delta \mathcal{L}_{\mathcal{N}_L(v)}(f_\theta), \quad (6)$$

174 with

$$\Delta \mathcal{L}_{\mathcal{N}_L(v)}(f_\theta) = \sum_{u \in \mathcal{N}_L(v)} m_u (\ell(f_\theta(\mathbf{X}, \mathbf{A}_{\mathcal{M}})_u, y_u) - \ell(f_\theta(\mathbf{X}, \mathbf{A}_{\tilde{\mathcal{M}}})_u, y_u)). \quad (7)$$

175 In (5), $\lambda = P(m_v = 1)$ denotes the prior probability that node v is a member of the training set,
 176 before observing the model, and σ is the sigmoid function. Also, the $n \times n$ matrices $\mathbf{A}_{\mathcal{M}}$ and $\mathbf{A}_{\tilde{\mathcal{M}}}$ in
 177 (6) and (7) are defined as

$$(\mathbf{A}_{\mathcal{M}})_{uw} = \begin{cases} A_{uw}, & m_u = m_w = 1, \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad (\mathbf{A}_{\tilde{\mathcal{M}}})_{uw} = \begin{cases} (\mathbf{A}_{\mathcal{M}})_{uw}, & u, w \neq v, \\ 0, & \text{otherwise} \end{cases}$$

178 for $u, w \in \mathcal{V}$.

179 *Proof.* See Appendix B.1. □

180 In (5), $S_{\mathcal{L}}(f_{\theta}, v, \tilde{\mathcal{M}}, \mathcal{G})$ represents the loss-based signal for the target node v , while the first loga-
 181 rithmic term corresponds to the expected signal over the distribution of the models induced by the
 182 membership configuration $\tilde{\mathcal{M}}$, $p(\phi|\tilde{\mathcal{M}}, \mathcal{G})$. The reference model distribution $p(\phi|\tilde{\mathcal{M}}, \mathcal{G})$ should
 183 be unbiased with respect to whether the target node is included in the training set [13]. The term
 184 $\Delta\mathcal{L}_{\mathcal{N}_L(v)}$ captures how the inclusion or exclusion of the target node influences the loss values of its
 185 neighbors—since an L -layer GNN aggregates information from nodes within the L -hop neighbor-
 186 hood, the optimal attack must consider this local graph structure around the target node.

187 The presence of the term (7) in the graph data setting, in contrast to the i.i.d. case (see [1, Thm. 2]
 188 and Corollary 1 in Section 5.1), highlights a key distinction: dependencies between data points play a
 189 central role.

190 To make predictions, we set a decision threshold τ such that target node v is inferred to be part of the
 191 training set if $P(m_v = 1|\theta, \mathcal{G}) > \tau$. For auditing purposes, it is important to evaluate performance
 192 across the full range of false positive rates, particularly at low false positive rate regimes, and we
 193 compute the receiver operating characteristic (ROC) curve by sweeping over $\tau \in [0, 1]$. Appendix C
 194 provides a detailed discussion on how an adversary can select the decision threshold τ .

195 4.3 G-BASE: Practical Bayes-Optimal MIA against GNNs

196 The Bayes-optimal membership inference score function in Theorem 1 is computationally intractable
 197 and thus necessitates approximations. In this section, we introduce a practical and effective approxi-
 198 mation to the Bayes-optimal decision rule, resulting in a powerful MIA.

199 The intractability stems from the two nested expectations in Equation (5): one over the membership
 200 statuses of all non-target samples, $\tilde{\mathcal{M}}$, and the other over the model distribution, $p(\phi|\tilde{\mathcal{M}}, \mathcal{G})$. The
 201 expectation over $\tilde{\mathcal{M}}$ requires sampling from the conditional distribution $P(\tilde{\mathcal{M}}|\theta, \mathcal{G})$, which is infea-
 202 sible since it is in general not available in closed form. To address this, we propose three approximate
 203 sampling strategies to generate samples from $P(\tilde{\mathcal{M}}|\theta, \mathcal{G})$: i) *model-independent sampling*, which
 204 assumes independence from the target model, i.e., $P(\tilde{\mathcal{M}}|\theta, \mathcal{G}) \approx P(\tilde{\mathcal{M}}|\mathcal{G})$; ii) *Markov chain Monte*
 205 *Carlo (MCMC) sampling*, which accounts explicitly for the dependence on the target model θ ; and
 206 iii) *0-hop MIA sampling*, which leverages a MIA attack for the i.i.d. setting to obtain per-node
 207 membership probabilities. These strategies are discussed in greater detail in Appendix B.2. Given
 208 any of these sampling strategies, we generate M samples $\tilde{\mathcal{M}}_1, \dots, \tilde{\mathcal{M}}_M$ and approximate the outer
 209 expectation $\mathbb{E}_{\tilde{\mathcal{M}} \sim P(\tilde{\mathcal{M}}|\theta, \mathcal{G})}[\cdot]$ in (5) with the sample average.

210 The remaining expectation over the model distribution $p(\phi|\tilde{\mathcal{M}}, \mathcal{G})$ (the integral term in (5)) is also
 211 intractable. We approximate it via Monte Carlo sampling using a set of shadow models trained on
 212 subgraphs of \mathcal{G} . We interpret the conditional distribution $p(\phi|\tilde{\mathcal{M}}, \mathcal{G})$ as the distribution of models
 213 trained on the subgraph defined by $\tilde{\mathcal{M}}$, and where the target node is included with probability 0.5. In
 214 a strict Bayesian sense, this models the uncertainty over whether the target node is part of the training
 215 set. However, directly sampling from this distribution would require training a separate set of shadow
 216 models for each pair $(v, \tilde{\mathcal{M}})$, which is computationally infeasible. For N target nodes, M samples of
 217 $\tilde{\mathcal{M}}$, and K shadow models per configuration, this would amount to training $N \times M \times K$ models.
 218 To reduce the number of shadow models, we approximate the posterior $p(\phi|\tilde{\mathcal{M}}, \mathcal{G})$ with its prior
 219 $p(\phi|\mathcal{G})$. This simplification allows us to reuse the same set of shadow models across all targets and
 220 sampled membership configurations $\tilde{\mathcal{M}}$. More precisely, the adversary samples subgraphs from \mathcal{G}
 221 and use them to train shadow models. Ideally, the adversary should sample subgraphs and train

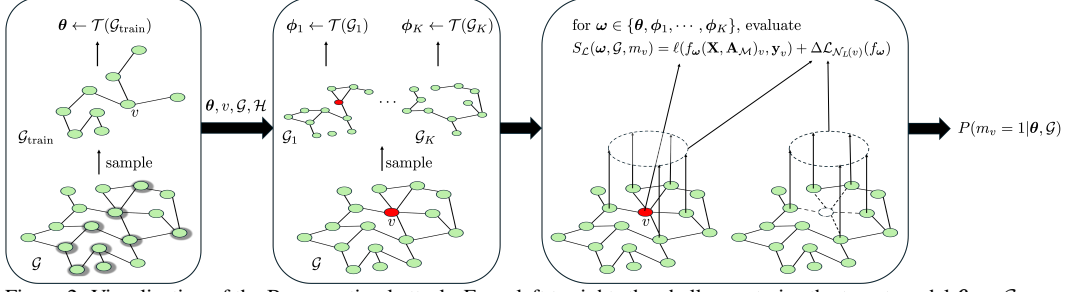


Figure 2: Visualization of the Bayes-optimal attack. From left to right: the challenger trains the target model θ on $\mathcal{G}_{\text{train}}$, and provides the trained model, a target node v , the underlying graph \mathcal{G} , and auxiliary information \mathcal{H} detailing the training procedure. The adversary then samples K graphs and trains a corresponding set of shadow models $\{\phi_i\}_{i=1}^K$. These sampled graphs may or may not contain the target node v . Finally, the adversary estimates the membership of v using the Bayes-optimal decision rule in (5), approximated via the Monte Carlo method in (8).

shadow models using a similar procedure as the challenger. However, this distribution is constrained by the adversary’s side information.

The integral in (5) can then be approximated as

$$\log \int e^{-S_{\mathcal{L}}(f_{\phi}, v, \tilde{\mathcal{M}}, \mathcal{G})} p(\phi | \tilde{\mathcal{M}}, \mathcal{G}) d\phi \approx \log \left(\frac{1}{K} \sum_{k=1}^K e^{-S_{\mathcal{L}}(\phi_k, \mathcal{G}, m_v)} \right), \quad \phi_k \leftarrow \mathcal{T}(\mathcal{G}_k), \quad (8)$$

where \mathcal{T} is the training algorithm.

Incorporating both approximations into (5), we arrive at the following attack:

Definition 2. (G-BASE attack) For a given threshold τ , and a set of shadow models $\{\phi_k\}_{k=1}^K$, target sample v is inferred to be part of the training set if $P(m_v = 1 | \theta, \mathcal{D}) > \tau$, where

$$P(m_v = 1 | \theta, \mathcal{G}) \approx \frac{1}{M} \sum_{i=1}^M \left[\sigma \left(-S_{\mathcal{L}}(f_{\theta}, v, \tilde{\mathcal{M}}, \mathcal{G}) - \log \left(\frac{1}{K} \sum_{k=1}^K e^{-S_{\mathcal{L}}(\phi_k, \mathcal{G}, m_v)} \right) + \log \frac{\lambda}{1 - \lambda} \right) \right]$$

We refer to this attack as the graph Bayes-approximate membership status estimation (G-BASE) attack. An illustration of the G-BASE attack is shown in Figure 2. The distinction between online and offline variants is discussed in Appendix F.

5 Bayes-Optimal MIAs for i.i.d. Data

In this section, we consider MIAs in the i.i.d. setting. We first show that the formulation in (5)–(6) recovers the result for the i.i.d. case presented in [1] (Theorem 2) as a special case. We then introduce a practical approximation of the Bayes-optimal decision rule that yields a powerful attack. The formal definition of the membership inference game in the i.i.d. setting is provided in Appendix D.

5.1 The Bayes-Optimal Decision Rule

The result below follows as a corollary of Theorem 1.

Corollary 1. Let $\mathcal{D} = \{(\mathbf{x}_v, \mathbf{y}_v)\}_{v=1}^n$ be a set of n i.i.d. data samples. For i.i.d. data, (5)–(6) reduces to

$$P(m_v = 1 | \theta, \mathcal{D}) = \mathbb{E}_{\tilde{\mathcal{M}} \sim P(\tilde{\mathcal{M}} | \theta, \mathcal{D})} \left[\sigma \left(-\ell(f_{\theta}(x_v), \mathbf{y}_v) - \log \int e^{-\ell(f_{\theta}(x_v), \mathbf{y}_v)} p(\phi | \tilde{\mathcal{M}}, \mathcal{D}) d\phi + \log \frac{\lambda}{1 - \lambda} \right) \right],$$

i.e., the result in [1, Thm. 2] with the temperature parameter absorbed into the loss function.

Proof. See Appendix E.1. \square

243 5.2 BASE: Practical Bayes-Optimal MIA for i.i.d. Data

244 As for the graph case, the Bayes-optimal decision rule stated in Theorem 1 and in [1, Thm. 2] (for
245 i.i.d. data) is computationally intractable. For the i.i.d. case, [1] proposed several approximations;
246 however, as shown in Section 6, the resulting attacks underperform compared to those in [11, 13]. In
247 this section, we introduce an alternative approximation to the Bayes-optimal decision rule for the
248 i.i.d. setting that yields a practical and powerful attack.

249 Due to space constraints, we present the formal definition of the proposed attack below and refer the
250 reader to Appendix E.2 for further details.

251 **Definition 3. (BASE attack)** For a given threshold τ , and a set of shadow models $\{\phi_k\}_{k=1}^K$, target
252 sample v is inferred to be part of the training set if $P(m_v = 1|\theta, \mathcal{D}) > \tau$, where

$$P(m_v|\theta, \mathcal{D}) = \sigma\left(-\ell(f_\theta(\mathbf{x}_v), \mathbf{y}_v) - \log\left(\frac{1}{K} \sum_{k=1}^K e^{-\ell(f_{\phi_k}(\mathbf{x}_v), \mathbf{y}_v)}\right) + \log \frac{\lambda}{1-\lambda}\right). \quad (9)$$

253 We refer to this attack as the Bayes-approximate membership status estimation (BASE) attack.

254 As shown in Section 6 and Appendix H, despite relying on a coarse approximation of the Bayes-
255 optimal decision rule in (5), the BASE attack matches or exceeds the performance of state-of-the-art
256 attacks [11, 13].

257 **Connection to RMIA.** We formally establish a connection between the BASE attack and the RMIA
258 attack proposed in [13]. To this end, we begin by introducing a notion of equivalence for score-based
259 MIAs.

260 **Definition 4. (MIA equivalence)** Two score-based MIAs are said to be equivalent if, for any decision
261 threshold for one attack, there exists a decision threshold for the other attack that yields identical
262 hard predictions.

263 In particular, equivalent attacks produce identical ROC curves.

264 The following theorem shows that the BASE attack is closely related to the RMIA attack [13].

265 **Theorem 2.** The BASE attack is equivalent to RMIA when $\gamma = 1$. More precisely, the membership
266 prediction scores produced by RMIA with $\gamma = 1$ is related to those of BASE via a monotone
267 increasing function.

268 *Proof.* See Appendix E.3. □

269 This result provides a theoretical justification for viewing RMIA as a Bayes-optimal attack, up to the
270 approximation on the shadow model distribution.

271 Despite the equivalence between BASE and RMIA (in the sense of Definition 4), BASE offers
272 significantly improved efficiency. In RMIA, the target model and shadow models need to be queried
273 for each target sample and each population sample. In a practical implementation, a sampled subset
274 $Z \sim \pi$ is typically used in place of π for efficiency, which can degrade performance [13]. BASE, on
275 the other hand, only needs to query the models using the target samples to match the performance of
276 RMIA with $Z = \pi$, which is also the best performing configuration of RMIA, see [13, Table 4].

278 6 Experiments

279 We evaluate the attack performance of BASE and G-BASE across a range of datasets and model
280 architectures. This section focuses on attacks against GNNs; results on i.i.d. data (CIFAR-10
281 and CIFAR-100) are presented in Appendix H.4. The code to reproduce our results is available at
282 <https://anonymous.4open.science/r/MIA-audit-GNN-7C32/>.

283 **Setup.** Experiments are conducted on 6 graph datasets: Cora, Citeseer, Pubmed, Flickr, Amazon-
284 Photo, and Github. Cora, Citeseer and Pubmed are citation networks previously used in node-level
285 MIA work [15–17]. For both the target and shadow models, we consider 2-layer GNN architectures:
286 GCN [28], GRAPHSAGE with max aggregation [32], and GAT [33]. Attack performance is measured
287 in terms of the area under the receiver operating characteristic curve (AUC), and true positive rate

(TPR) at 1% and 0.1% false positive rate (FPR). The AUC captures the average attack performance across all FPRs, while the attack performance at low FPR is most relevant in practice, as high TPR at a low FPR is essential for confident and reliable membership inference [11]. Additional details of the training procedure and hyperparameter selection are provided in Appendix A.

Baseline attacks. We compare the performance of BASE and G-BASE against LiRA [11] and RMIA [13], the current state-of-the-art MIAs for i.i.d. data. We also compare against node-level MLP-classifier attacks against GNNs from prior work [15–17]. To strengthen these baselines, we enhance the MLP-classifier attacks by using multiple shadow models to generate attack features, following the shadow training framework of [2]. Additional details on the MLP-classifier attacks are provided in Appendix A. The attacks are implemented based on the descriptions in their respective papers, incorporating information from the authors’ code when available.

Online and offline setting. We evaluate both online and offline attacks, as defined in Appendix F, and compare BASE and G-BASE with LiRA and RMIA in both settings. MLP-classifier attacks are considered only in the online setting, since shadow models are only used to construct features for the attack model and not to estimate model distributions.

Attacks against GNNs. Figure 1 shows the ROC for G-BASE (using model-independent sampling) and several baselines on Flickr, using a 2-layer GCN architecture and 8 shadow models. G-BASE consistently outperforms all baselines across the full range of FPRs. Table 1 shows the performance of BASE and G-BASE and the baseline MIAs across multiple datasets and GNN target models. All attacks use the same set of K shadow models for a fair comparison. The number of sampled graphs in G-BASE is set to 8 when 4 or 8 shadow models are used, and 32 when 64 or 128 shadow models are used, except for Flickr, where 16 sampled graphs are used for large K . We evaluate G-BASE for both model-independent sampling (MI) and 0-hop MIA sampling (MIA) (see Section 4.3), using BASE as the 0-hop MIA. In offline mode, only the best-performing sampling strategy is reported. Results with MCMC sampling are presented in Appendix H.3. The MLP-classifier attack is evaluated in two variants: using 0-hop query outputs as attack features, and using concatenated 0-hop and 2-hop query outputs to form attack features (0 + 2-hop), with 50% edge dropout in the latter case. The hyperparameters for offline RMIA and offline BASE are selected using Bayesian optimization [34], with one shadow model acting as a simulated target model. For offline G-BASE, we use $\alpha = \{0.9, 1\}$, which we recommend as defaults, noting that performance is largely insensitive to this hyperparameter.

We make the following key observations: The MLP-classifier attacks are not competitive with our attacks BASE and G-BASE, or LiRA and RMIA across all datasets and target models and perform as random guessing on the larger, more challenging datasets. For $K = 4$ and 8 shadow models, BASE and G-BASE consistently outperform all baselines on the larger and more challenging datasets Pubmed, Flickr, Amazon-Photo, and Github. On these datasets, G-BASE effectively leverages local neighborhood information of the target node to enhance the attack performance. On the smaller datasets Cora and Citeseer (≈ 3000 nodes), LiRA achieves comparatively good performance in terms of AUC. However, BASE and G-BASE achieve superior performance in the more relevant low-FPR regime. As predicted by Theorem 2, in the online setting, BASE and RMIA yield nearly identical results. Minor differences stem from RMIA using only half of the nodes in the Z set; including all nodes in the Z set, makes the attacks equivalent, as established in Theorem 2. However, as noted in 5.2, BASE requires significantly less computation in terms of model queries, compared to RMIA. In fact, BASE requires only a single query of the target and shadow models, whereas RMIA also need to query the models over the Z set. In the case of a large number of shadow models, the performance of LiRA increases significantly. On some datasets, most notably Cora and Citeseer, LiRA performs very well. However, on other datasets, e.g. Flickr and Pubmed, the performance gain of LiRA is more modest. It has to do with how well justified the normality assumption of LiRA is for the given dataset and model. Our attacks BASE and G-BASE are more principled and robust against datasets and models, and achieve top performance in many settings and also for a large number of shadow models K . However, we consider attacks that rely on a large number of shadow model to be proof-of-concept, since training a large number of shadow models is often infeasible for large real-world datasets.

Robustness to mismatched adversary assumptions. We evaluate and compare the attacks in a more challenging setting where the adversary lacks precise knowledge of the challenger’s training procedure and model architecture. Specifically, the challenger trains

Table 1: Comparison of different attacks across datasets and model architectures. Performance is measured in terms of AUC and TPR at 1% and 0.1% FPR. The result is reported as the sample mean plus-minus the (population) standard deviation over 10 random target models and samples of target nodes. The parameter K denotes the number of shadow models. Our attacks BASE and G-BASE achieves top performance in the majority of cases, sometimes with a significant margin.

| K | ATTACK | CORA (GCN) | | | CITSEER (GAT) | | | PUBMED (GRAPHSAGE) | | |
|-----|---------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|-----------------------------------|-----------------------------------|
| | | AUC (%) | TPR@FPR (%) | | AUC (%) | TPR@FPR (%) | | AUC (%) | TPR@FPR (%) | |
| | | | 1% | 0.1% | | 1% | 0.1% | | 1% | 0.1% |
| 8 | MLP (0-HOP) | 69.11 \pm 2.03 | 3.13 \pm 0.72 | 0.50 \pm 0.60 | 73.17 \pm 0.99 | 5.15 \pm 1.32 | 0.73 \pm 0.45 | 49.89 \pm 0.70 | 0.97 \pm 0.21 | 0.09 \pm 0.05 |
| | MLP (0+2-HOP) | 64.23 \pm 1.29 | 3.31 \pm 1.07 | 0.84 \pm 0.56 | 65.52 \pm 1.39 | 3.02 \pm 0.88 | 0.46 \pm 0.46 | 50.52 \pm 0.46 | 1.10 \pm 0.17 | 0.11 \pm 0.07 |
| | LIRA | 81.86 \pm 0.62 | 8.21 \pm 3.46 | 1.15 \pm 1.01 | 85.33 \pm 0.41 | 15.69 \pm 2.99 | 2.55 \pm 3.71 | 53.08 \pm 0.58 | 1.16 \pm 0.14 | 0.10 \pm 0.07 |
| | RMIA | 81.48 \pm 1.46 | 15.63 \pm 2.13 | 3.77 \pm 2.47 | 83.94 \pm 0.64 | 20.02 \pm 2.95 | 5.11 \pm 3.56 | 57.20 \pm 0.66 | 2.85 \pm 0.36 | 0.45 \pm 0.12 |
| | BASE | 81.48 \pm 1.46 | 15.66 \pm 2.15 | 3.77 \pm 2.47 | 83.94 \pm 0.64 | 20.04 \pm 2.94 | 5.11 \pm 3.56 | 57.20 \pm 0.66 | 2.85 \pm 0.35 | 0.45 \pm 0.12 |
| | G-BASE (MI) | 77.37 \pm 0.82 | 8.42 \pm 2.01 | 0.93 \pm 0.78 | 81.02 \pm 0.96 | 11.32 \pm 2.49 | 1.41 \pm 1.62 | 62.96 \pm 0.42 | 5.22 \pm 0.50 | 1.07 \pm 0.51 |
| | G-BASE (MIA) | 77.38 \pm 1.02 | 15.55 \pm 2.20 | 5.83 \pm 2.94 | 82.21 \pm 0.84 | 20.25 \pm 2.15 | 5.99 \pm 3.62 | 62.95 \pm 0.52 | 5.24 \pm 0.43 | 1.16 \pm 0.29 |
| | LIRA (OFF) | 81.81 \pm 0.92 | 10.47 \pm 4.33 | 2.47 \pm 3.05 | 84.07 \pm 0.74 | 19.48 \pm 3.30 | 4.03 \pm 3.02 | 55.27 \pm 0.57 | 1.35 \pm 0.25 | 0.15 \pm 0.08 |
| 4 | RMIA (OFF) | 79.67 \pm 1.23 | 15.57 \pm 2.30 | 6.91 \pm 3.60 | 82.46 \pm 0.77 | 23.63 \pm 2.31 | 8.11 \pm 2.73 | 56.51 \pm 0.65 | 3.04 \pm 0.27 | 0.75 \pm 0.18 |
| | BASE (OFF) | 80.53 \pm 0.90 | 16.50 \pm 2.36 | 6.35 \pm 3.95 | 84.03 \pm 0.77 | 23.20 \pm 2.36 | 8.99 \pm 3.93 | 56.72 \pm 0.69 | 3.08 \pm 0.26 | 0.76 \pm 0.19 |
| | G-BASE (OFF) | 74.87 \pm 1.51 | 9.70 \pm 1.63 | 3.16 \pm 1.19 | 77.16 \pm 1.29 | 14.96 \pm 1.97 | 4.58 \pm 1.43 | 62.68 \pm 0.48 | 5.34 \pm 0.37 | 1.32 \pm 0.37 |
| 128 | MLP (0-HOP) | 72.28 \pm 1.78 | 4.37 \pm 1.57 | 1.21 \pm 0.79 | 74.28 \pm 1.32 | 5.29 \pm 1.86 | 1.03 \pm 0.95 | 50.83 \pm 0.83 | 1.06 \pm 0.35 | 0.11 \pm 0.09 |
| | MLP (0+2-HOP) | 66.86 \pm 1.52 | 2.70 \pm 1.10 | 0.59 \pm 0.50 | 68.60 \pm 1.72 | 3.86 \pm 1.12 | 0.84 \pm 0.72 | 51.06 \pm 1.12 | 1.16 \pm 0.22 | 0.10 \pm 0.07 |
| | LIRA | 89.23 \pm 0.73 | 35.52 \pm 4.83 | 20.30 \pm 4.80 | 91.60 \pm 0.53 | 47.50 \pm 2.13 | 27.87 \pm 5.08 | 56.58 \pm 0.94 | 3.37 \pm 0.50 | 1.12 \pm 0.44 |
| | RMIA | 82.70 \pm 1.02 | 19.79 \pm 3.07 | 6.79 \pm 5.28 | 84.95 \pm 0.83 | 24.00 \pm 2.40 | 7.21 \pm 4.46 | 57.47 \pm 0.78 | 3.35 \pm 0.54 | 0.65 \pm 0.23 |
| | BASE | 82.70 \pm 1.02 | 19.79 \pm 3.07 | 6.79 \pm 5.28 | 84.95 \pm 0.83 | 24.00 \pm 2.40 | 7.21 \pm 4.46 | 57.47 \pm 0.78 | 3.35 \pm 0.54 | 0.65 \pm 0.23 |
| | G-BASE | 78.42 \pm 1.01 | 19.16 \pm 2.89 | 9.04 \pm 3.34 | 83.48 \pm 1.02 | 26.04 \pm 2.97 | 10.28 \pm 4.15 | 63.11 \pm 1.05 | 5.57 \pm 0.92 | 1.39 \pm 0.61 |
| | LIRA (OFF) | 85.98 \pm 1.02 | 33.50 \pm 3.79 | 19.73 \pm 4.35 | 87.32 \pm 0.61 | 45.40 \pm 1.77 | 25.90 \pm 5.40 | 56.07 \pm 0.98 | 2.38 \pm 0.56 | 0.42 \pm 0.19 |
| | RMIA (OFF) | 80.80 \pm 1.47 | 19.85 \pm 3.23 | 12.64 \pm 3.61 | 83.92 \pm 0.76 | 30.31 \pm 3.02 | 15.42 \pm 4.07 | 57.27 \pm 0.90 | 3.36 \pm 0.51 | 0.92 \pm 0.33 |
| 64 | BASE (OFF) | 81.50 \pm 1.63 | 22.05 \pm 4.20 | 13.23 \pm 3.70 | 85.45 \pm 0.76 | 31.19 \pm 3.47 | 12.17 \pm 4.12 | 57.36 \pm 1.04 | 3.38 \pm 0.50 | 0.91 \pm 0.30 |
| | G-BASE (OFF) | 77.17 \pm 1.04 | 13.88 \pm 2.55 | 5.67 \pm 3.09 | 80.17 \pm 0.86 | 21.47 \pm 2.05 | 9.12 \pm 2.47 | 63.77 \pm 0.87 | 6.50 \pm 0.76 | 2.15 \pm 0.93 |

| K | ATTACK | FLICKR (GCN) | | | AMAZON-PHOTO (GAT) | | | GITHUB (GRAPHSAGE) | | |
|-----|---------------|------------------------------------|-----------------------------------|-----------------------------------|------------------------------------|-----------------------------------|-----------------------------------|------------------------------------|-----------------------------------|-----------------------------------|
| | | AUC (%) | TPR@FPR (%) | | AUC (%) | TPR@FPR (%) | | AUC (%) | TPR@FPR (%) | |
| | | | 1% | 0.1% | | 1% | 0.1% | | 1% | 0.1% |
| 8 | MLP (0-HOP) | 50.04 \pm 0.32 | 1.03 \pm 0.14 | 0.13 \pm 0.07 | 51.65 \pm 0.95 | 1.40 \pm 0.38 | 0.26 \pm 0.15 | 49.97 \pm 0.41 | 1.01 \pm 0.15 | 0.10 \pm 0.03 |
| | MLP (0+2-HOP) | 49.92 \pm 0.41 | 1.04 \pm 0.15 | 0.10 \pm 0.03 | 50.69 \pm 1.10 | 1.10 \pm 0.24 | 0.09 \pm 0.08 | 50.18 \pm 0.36 | 1.10 \pm 0.17 | 0.11 \pm 0.03 |
| | LIRA | 51.41 \pm 0.44 | 1.07 \pm 0.11 | 0.08 \pm 0.03 | 54.92 \pm 1.11 | 2.28 \pm 0.56 | 0.27 \pm 0.31 | 51.17 \pm 0.41 | 1.00 \pm 0.17 | 0.12 \pm 0.04 |
| | RMIA | 56.22 \pm 0.68 | 1.97 \pm 0.35 | 0.24 \pm 0.11 | 56.64 \pm 0.87 | 2.35 \pm 0.67 | 0.44 \pm 0.37 | 54.52 \pm 0.49 | 2.08 \pm 0.29 | 0.37 \pm 0.13 |
| | BASE | 56.22 \pm 0.68 | 1.97 \pm 0.35 | 0.24 \pm 0.11 | 56.64 \pm 0.87 | 2.36 \pm 0.67 | 0.44 \pm 0.37 | 54.52 \pm 0.49 | 2.08 \pm 0.30 | 0.37 \pm 0.13 |
| | G-BASE (MI) | 60.46 \pm 0.94 | 2.97 \pm 0.33 | 0.46 \pm 0.18 | 56.52 \pm 0.61 | 3.94 \pm 0.73 | 0.61 \pm 0.30 | 57.71 \pm 0.93 | 2.83 \pm 0.28 | 0.57 \pm 0.16 |
| | G-BASE (MIA) | 57.67 \pm 0.62 | 2.52 \pm 0.23 | 0.41 \pm 0.10 | 56.74 \pm 0.47 | 3.93 \pm 0.99 | 0.89 \pm 0.54 | 57.76 \pm 0.67 | 2.96 \pm 0.28 | 0.65 \pm 0.17 |
| | LIRA (OFF) | 55.78 \pm 0.62 | 1.59 \pm 0.20 | 0.15 \pm 0.06 | 57.12 \pm 1.08 | 2.96 \pm 0.46 | 0.50 \pm 0.30 | 53.50 \pm 0.40 | 1.28 \pm 0.14 | 0.14 \pm 0.05 |
| 4 | RMIA (OFF) | 56.12 \pm 0.64 | 2.00 \pm 0.17 | 0.27 \pm 0.06 | 56.48 \pm 0.80 | 3.55 \pm 0.71 | 1.32 \pm 0.56 | 54.28 \pm 0.49 | 2.09 \pm 0.34 | 0.36 \pm 0.12 |
| | BASE (OFF) | 56.13 \pm 0.64 | 1.99 \pm 0.22 | 0.23 \pm 0.06 | 56.63 \pm 1.00 | 3.70 \pm 0.74 | 1.09 \pm 0.58 | 54.25 \pm 0.49 | 2.12 \pm 0.35 | 0.40 \pm 0.11 |
| | G-BASE (OFF) | 60.40 \pm 0.98 | 2.97 \pm 0.51 | 0.51 \pm 0.12 | 57.63 \pm 1.06 | 4.77 \pm 0.46 | 1.28 \pm 0.45 | 58.38 \pm 0.61 | 3.12 \pm 0.50 | 0.63 \pm 0.19 |
| 128 | MLP (0-HOP) | 50.36 \pm 0.69 | 1.12 \pm 0.26 | 0.14 \pm 0.05 | 54.33 \pm 1.15 | 1.87 \pm 0.28 | 0.50 \pm 0.13 | 50.06 \pm 1.00 | 1.01 \pm 0.28 | 0.13 \pm 0.08 |
| | MLP (0+2-HOP) | 50.50 \pm 0.66 | 1.11 \pm 0.17 | 0.14 \pm 0.07 | 51.33 \pm 1.03 | 1.11 \pm 0.41 | 0.19 \pm 0.10 | 50.06 \pm 0.61 | 0.98 \pm 0.20 | 0.15 \pm 0.08 |
| | LIRA | 55.02 \pm 0.50 | 1.97 \pm 0.22 | 0.31 \pm 0.12 | 58.79 \pm 0.63 | 6.07 \pm 0.81 | 3.00 \pm 0.48 | 53.23 \pm 0.88 | 1.92 \pm 0.27 | 0.30 \pm 0.12 |
| | RMIA | 56.03 \pm 0.49 | 1.97 \pm 0.26 | 0.33 \pm 0.10 | 56.27 \pm 0.90 | 2.49 \pm 0.40 | 0.28 \pm 0.20 | 54.56 \pm 0.93 | 2.10 \pm 0.33 | 0.34 \pm 0.12 |
| | BASE | 56.03 \pm 0.49 | 1.97 \pm 0.26 | 0.33 \pm 0.10 | 56.27 \pm 0.90 | 2.49 \pm 0.40 | 0.28 \pm 0.20 | 54.56 \pm 0.93 | 2.10 \pm 0.33 | 0.34 \pm 0.12 |
| | G-BASE | 59.87 \pm 1.65 | 3.07 \pm 0.46 | 0.51 \pm 0.22 | 55.10 \pm 0.80 | 4.12 \pm 0.77 | 1.12 \pm 0.35 | 58.02 \pm 0.97 | 3.04 \pm 0.28 | 0.59 \pm 0.07 |
| | LIRA (OFF) | 56.19 \pm 0.54 | 1.89 \pm 0.28 | 0.36 \pm 0.09 | 57.47 \pm 0.55 | 5.50 \pm 0.67 | 2.21 \pm 0.55 | 54.00 \pm 1.00 | 1.56 \pm 0.29 | 0.20 \pm 0.13 |
| | RMIA (OFF) | 56.14 \pm 0.50 | 2.19 \pm 0.43 | 0.39 \pm 0.15 | 55.97 \pm 1.16 | 4.00 \pm 0.31 | 1.08 \pm 0.42 | 54.62 \pm 0.92 | 2.45 \pm 0.30 | 0.44 \pm 0.12 |
| 64 | BASE (OFF) | 56.01 \pm 0.52 | 2.19 \pm 0.38 | 0.43 \pm 0.19 | 56.46 \pm 1.07 | 3.96 \pm 0.35 | 1.03 \pm 0.39 | 54.64 \pm 0.93 | 2.51 \pm 0.31 | 0.45 \pm 0.12 |
| | G-BASE (OFF) | 60.59 \pm 1.44 | 3.03 \pm 0.67 | 0.49 \pm 0.16 | 57.10 \pm 1.08 | 5.48 \pm 0.81 | 2.11 \pm 0.51 | 58.91 \pm 0.67 | 3.45 \pm 0.64 | 0.73 \pm 0.22 |

a 2-layer GCN on 35% of Cora, while the adversary trains 8 2-layer GAT shadow models on 50% of the nodes, following the shadow model training procedure outlined in Appendix A. Furthermore, the challenger trains the shadow model using a SGD optimizer with momentum, while the adversary uses an Adam optimizer.

Figure 3 presents the results of our attacks and baseline attacks (see Appendix H.2 for more results). In this setting, all attacks show degraded performance, with AUC scores dropping by 10 percentage points or more across all attacks, compared to the performance against a GCN target model on Cora (see Table 1). LIRA suffers the most, losing 20 AUC percentage points and performing worse than MLP-classifier attacks in the low FPR regime. In contrast, our attacks and RMIA are more robust to mismatches in model and training procedures introduced by the adversary, with our offline BASE attack achieving the highest TPR at low FPRs.

Conclusions. We proposed BASE and G-BASE, practical and theoretically-grounded MIAs for i.i.d. and graph data. By deriving the Bayes-optimal inference rule for node-level attacks on GNNs, we addressed key challenges posed by structural dependencies in graphs. Our attacks match or surpass existing state-of-the-art methods (LIRA and RMIA) while requiring significantly lower computational overhead. Our results bridge the gap between theoretical optimality and practical implementation of MIAs, offering an efficient framework for privacy auditing across both classical and graph-based learning settings.

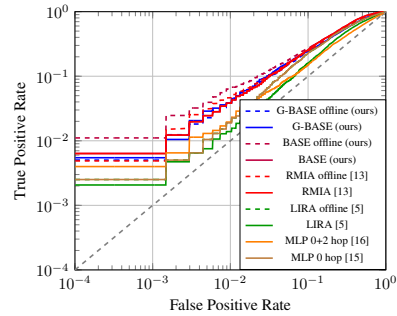


Figure 3: ROC of a mismatched attack averaged over 10 independent target models. Shadow models utilize a GAT architecture and different training procedure. 8 shadow models for online; 4 for offline.

References

- [1] Alexandre Sablayrolles, Matthijs Douze, Yann Ollivier, Cordelia Schmid, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *ICML*, 2019.
- [2] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (SP)*, 2017.
- [3] Jiayuan Ye, Anastasia Borovykh, Soufiane Hayou, and Reza Shokri. Leave-one-out distinguishability in machine learning. In *ICLR*, 2024.
- [4] Awni Hannun, Chuan Guo, and Laurens van der Maaten. Measuring data leakage in machine-learning models with fisher information. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2021.
- [5] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *USENIX Security*, August 2021.
- [6] European Data Protection Board (EDPB). Edpb opinion 2024/28 on ai models. Technical report, European Data Protection Board, December 2024.
- [7] Jie Zhang, Debeshee Das, Gautam Kamath, and Florian Tramèr. Membership inference attacks cannot prove that a model was trained on your data. *arXiv preprint arXiv:2409.19798*, 2024.
- [8] Milad Nasr, Shuang Songi, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlin. Adversary instantiation: Lower bounds for differentially private machine learning. In *IEEE Symposium on security and privacy (SP)*, 2021.
- [9] Martin Bertran, Shuai Tang, Aaron Roth, Michael Kearns, Jamie H Morgenstern, and Steven Z Wu. Scalable membership inference attacks via quantile regression. *Neurips*, 36, 2023.
- [10] Hao Li, Zheng Li, Siyuan Wu, Chengrui Hu, Yutong Ye, Min Zhang, Dengguo Feng, and Yang Zhang. SeqMIA: Sequential-metric based membership inference attack. In *ACM CCS*, 2024.
- [11] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy (SP)*, 2022.
- [12] Yiyong Liu, Zhengyu Zhao, Michael Backes, and Yang Zhang. Membership inference attacks by exploiting loss trajectory. In *ACM CCS*, 2022.
- [13] Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks. In *ICML*, 2024.
- [14] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *ACM CCS*, 2022.
- [15] Iyiola E Olatunji, Wolfgang Nejdl, and Megha Khosla. Membership inference attack on graph neural networks. In *IEEE TPS-ISA*, 2021.
- [16] Vasisht Duddu, Antoine Boutet, and Virat Shejwalkar. Quantifying privacy leakage in graph embedding. In *International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous)*, 2020.
- [17] Xinlei He, Rui Wen, Yixin Wu, Michael Backes, Yun Shen, and Yang Zhang. Node-level membership inference attacks against graph neural networks. *arXiv:2102.05429 [cs.CR]*, 2021.
- [18] Mauro Conti, Jiaxin Li, Stjepan Picek, and Jing Xu. Label-only membership inference attack against node-level graph neural networks. In *AISeC*, 2022.
- [19] Abdellah Jnaini, Afafe Bettar, and Mohammed Amine Koulali. How powerful are membership inference attacks on graph neural networks? In *SSDBM*, 2022.

- 412 [20] Fan Wu, Yunhui Long, Ce Zhang, and Bo Li. Linkteller: Recovering private edges from graph
413 neural networks via influence analysis. In *IEEE SP*, 2022.
- 414 [21] Xinlei He, Jinyuan Jia, Michael Backes, Neil Zhenqiang Gong, and Yang Zhang. Stealing links
415 from graph neural networks. In *USENIX Security*, 2020.
- 416 [22] Xiuling Wang and Wendy Hui Wang. Link membership inference attacks against unsupervised
417 graph representation learning. In *ACSAC*, 2023.
- 418 [23] Ruyi Ding, Shijin Duan, Xiaolin Xu, and Yunsu Fei. Vertexserum: Poisoning graph neural
419 networks for link inference. In *ICCV*, 2023.
- 420 [24] Kailai Li, Jiawei Sun, Ruoxin Chen, Wei Ding, Kexue Yu, Jie Li, and Chentao Wu. Towards
421 practical edge inference attacks against graph neural networks. In *ICASSP*, 2023.
- 422 [25] Bang Wu, Xiangwen Yang, Shirui Pan, and Xingliang Yuan. Adapting membership inference
423 attacks to gnn for graph classification: Approaches and implications. In *ICDM*, 2021.
- 424 [26] Junze Yang, Hongwei Li, Wenshu Fan, Xilin Zhang, and Meng Hao. Membership inference
425 attacks against the graph classification. In *GLOBECOM*, 2023.
- 426 [27] Fabian P. Krüger, Johan Östman, Lewis Mervin, Igor V. Tetko, and Ola Engkvist. Publishing
427 neural networks in drug discovery might compromise training data privacy. *J. Cheminf.*, 2025.
- 428 [28] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional
429 networks, 2017.
- 430 [29] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally
431 connected networks on graphs, 2014. URL <https://arxiv.org/abs/1312.6203>.
- 432 [30] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks
433 on graphs with fast localized spectral filtering, 2017. URL <https://arxiv.org/abs/1606.09375>.
- 434
- 435 [31] William L. Hamilton. Graph representation learning. *Synthesis Lectures on Artificial Intelligence
436 and Machine Learning*, 14(3):1–159, 2020.
- 437 [32] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large
438 graphs, 2017.
- 439 [33] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua
440 Bengio. Graph attention networks. In *ICLR*, 2018.
- 441 [34] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna:
442 A next-generation hyperparameter optimization framework. In *ACM SIGKDD International
443 Conference on Knowledge Discovery & Data Mining*, 2019.
- 444 [35] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*,
445 2015.
- 446 [36] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*.
447 MIT Press, 2009.
- 448 [37] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint
449 arXiv:1605.07146*, 2017.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: We show that our attacks are competitive with previous state-of-the-art attacks in Section 6. The asserted connection of our attack and the prior state-of-the-art attack RMIA is proven in Section 5.2.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: The paper explicitly discusses limitations in terms of the datasets in which the different methods are compared and selecting the decision threshold when the adversary does not have access to the ground truth.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: The set of assumptions is stated within each theorem. We provide complete and carefully checked proofs of the theorems in the appendices provided in the supplemental material. All the theorems, lemmas, formulas, and proofs are numbered and cross-referenced.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We provide a clear description of the experimental setup in Section 6 of the main paper and Appendix A of the supplemental material. We provide a link to a repository containing the code to reproduce our results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide a link to a repository containing the code to reproduce our results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We specify all the training and test details, including the datasets, model architecture, performance metric, hyper-parameter tuning.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report the ROC curves and extract the area under the curve, as well as the true positive rate at a low false positive rate. These are suitable and typically used metrics in all existing membership inference attacks.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: The paper currently does not include explicit details about the computational resources required; this will be provided in the supplemental material upon acceptance. The paper introduces a new framework and computational resource details are not crucial for understanding or replicating the main contributions and their impact.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: The paper complies fully with the NeurIPS Code of Ethics, involving no ethical concerns or misuse.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses positive societal impacts by enhancing privacy and classical and graph-based learning scenarios. Although we design attacks, the purpose is to understand the vulnerability of trained models, which eventually assists in the design of privacy-enhancing mechanisms. No negative societal impacts were identified.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not involve datasets or models with high risks for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets and existing models used are clearly cited with references and licenses properly respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA] .

Justification: The paper does not introduce any new datasets, models, or code assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA] .

Justification: The research presented in this paper does not involve crowdsourcing or human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The research does not involve human subjects and thus does not require IRB approvals.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The research does not utilize large language models as part of its core methodology.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Experimental Setup

Target Model Training. We consider target models trained in a supervised manner for node classification tasks, using the commonly adopted cross-entropy loss as the objective function. We evaluate three target model architectures: graph convolutional networks (GCNs) [28], GraphSAGE [32] with max aggregation, and graph attention networks (GATs) [33] using 4 attention heads in the first layer and 2 in the second. Optimization is performed using Adam [35]. For each dataset and model, hyperparameters are selected via a grid search, including the learning rate, weight decay, number of training epochs, dropout rate, and dimension of the first GNN layer.¹ In particular, we search over $\{0.01, 0.001\}$ for the learning rate, $\{0.0001, 0.00001\}$ for the weight decay, and $\{0.0, 0.25, 0.5\}$ for the dropout rate. For the hidden dimension of the first layer, we search in $\{32, 64, 128, 256, 512\}$, with 32 or 512 excluded depending on the dataset. The initial search space for the number of epochs is typically $\{20, 50, 100, 200, 400, 800, 1600\}$, and is sometimes later refined. Target models with very small generalization gaps are often difficult to attack, making it harder to discern differences in MIA performance. To avoid this issue, we restrict the grid search to configurations that yield an average generalization gap of at least 8%. This represents a realistic generalization gap, sufficient to enable meaningful attacks without excessive overfitting that would distort expected attack performance. We emphasize that our goal is not to produce the best performing target model, but rather to obtain well-performing target models with a representative generalization gap.

The dataset-model combinations evaluated in Table 1 correspond to the best-performing model architecture selected for each dataset. The corresponding train and test accuracies of the target models are reported in Table 2.

Table 2: Train and test accuracy of the target model on the different datasets and architectures used in Table 1. The accuracies are reported as mean \pm (population) standard deviation, over the 10 different target models used to evaluate the 8/4 shadow model attacks. The train and test accuracies for the 128/64 shadow model attack evaluations are similar.

| Dataset (model) | Train accuracy | Test accuracy |
|--------------------|---------------------|---------------------|
| Cora (GCN) | 0.9603 ± 0.0042 | 0.8117 ± 0.0139 |
| Citeseer (GAT) | 0.9221 ± 0.0064 | 0.7411 ± 0.0143 |
| Pubmed (GraphSAGE) | 0.9629 ± 0.0021 | 0.8727 ± 0.0027 |
| Flickr (GCN) | 0.5749 ± 0.0104 | 0.4741 ± 0.0043 |
| Amazon-photo (GAT) | 0.9970 ± 0.0009 | 0.9100 ± 0.0048 |
| Github (GraphSAGE) | 0.9376 ± 0.0147 | 0.8442 ± 0.0066 |

Shadow model training. Shadow models are trained using the same hyperparameter settings as the target model, implicitly assuming adversarial side-knowledge. This setting is particularly relevant for MIA auditing, as it yields an upper bound on the attack performance. To facilitate efficient MIA auditing in the online setting (see Appendix F for a discussion of online vs. offline settings), we adopt the shadow model training procedure proposed in [11] and also used in [13]. Specifically, each shadow model is trained on half of the data population (e.g., half the nodes in a graph dataset), such that each data sample is included in the training set of half of the models. Pseudo-code for our precise shadow model training procedure is provided in Algorithm 1. For graph data, the data population is a graph dataset, and sampling data points corresponds to sampling nodes, retaining the edges between sampled nodes. This procedure guarantees a balanced set of in-models (models trained with the target data point) and out-models (models trained without it) for each data point. In the offline setting, for each target sample, the in-models for that sample are filtered out, so that only out-models are used in the attack. This filtering approach eliminates the need for a separate, disjoint dataset to train shadow models, an important advantage in graph-data settings, since splitting a graph in disjoint parts reduces the number of edges, leading to sparse graphs when the full graph has low average degree. The downside of the filtering approach is that only half of the shadow models are used for attacking any given target sample.

¹The second layer always produces embeddings with dimensionality equal to the number of classes.

Algorithm 1 Shadow Model Training Procedure.

```

1: Input: Data population  $\mathcal{G}$ , training algorithm  $\mathcal{T}$ , and even number of shadow models  $2N$ .
2:  $\Phi \leftarrow \emptyset$ 
3: for  $k = 1$  to  $N$  do
4:    $\mathcal{G}_k \sim \text{Uniform}(\mathcal{G}), |\mathcal{G}_k| = \frac{1}{2}|\mathcal{G}|$ 
5:    $\mathcal{G}_k^c = \{z : z \in \mathcal{G}, z \notin \mathcal{G}_k\}$ 
6:    $\phi_k \leftarrow \mathcal{T}(\mathcal{G}_k)$ 
7:    $\phi_k^c \leftarrow \mathcal{T}(\mathcal{G}_k^c)$ 
8:    $\Phi \leftarrow \Phi \cup \{\phi_k, \phi_k^c\}$ 
9: end for
10: return  $\Phi$ 

```

829 B Bayes-Optimal Node-Level MIAs Against GNNs

830 B.1 Proof of Theorem 1

831 The proof follows along the lines of the proofs of [1, Thms. 1 and 2] for the case of i.i.d. data. We
832 begin by applying the law of total expectation to express $P(m_v = 1|\theta, \mathcal{G})$ as an expectation over the
833 unknown membership statuses of the remaining nodes,

$$P(m_v = 1|\theta, \mathcal{G}) = \mathbb{E}_{\tilde{\mathcal{M}} \sim P(\tilde{\mathcal{M}}|\theta, \mathcal{G})}[P(m_v = 1|\tilde{\mathcal{M}}, \theta, \mathcal{G})]. \quad (10)$$

834 Unlike [1], where the expectation is taken over the data samples, our threat model assumes that
835 the full graph \mathcal{G} is known to the adversary. Consequently, we marginalize only over the unknown
836 membership indicator variables of the non-target nodes.

837 The term $P(m_v = 1|\tilde{\mathcal{M}}, \theta, \mathcal{G})$ can be computed using Bayes' rule, yielding

$$P(m_v = 1|\theta, \mathcal{G}) = \mathbb{E}_{\tilde{\mathcal{M}} \sim P(\tilde{\mathcal{M}}|\theta, \mathcal{G})} \left[\sigma \left(\log \frac{p(\theta|m_v = 1, \tilde{\mathcal{M}}, \mathcal{G})}{p(\theta|m_v = 0, \tilde{\mathcal{M}}, \mathcal{G})} + \log \frac{\lambda}{1 - \lambda} \right) \right], \quad (11)$$

838 where $\lambda = P(m_v = 1|\tilde{\mathcal{M}}, \mathcal{G})$ is the prior probability that node v is a member of the training
839 set, before observing the model. Under our threat model, conditioning only on the data and the
840 membership indicators of other nodes does not provide any information about the target node
841 membership status. Hence, $\lambda = P(m_v = 1)$.

842 Finally, we need to compute the log ratio of model posteriors in (11). Assuming the negative log-
843 likelihood loss function defined in (3), the posterior distribution of the model parameters can be
844 expressed using Bayes' rule in terms of the loss function and a prior $p(\theta|\mathcal{M}, \mathbf{X}, \mathbf{A})$ as

$$\begin{aligned}
p(\theta|\mathcal{M}, \mathcal{G}) &= \frac{p(\mathbf{Y}|\theta, \mathcal{M}, \mathbf{X}, \mathbf{A})p(\theta|\mathcal{M}, \mathbf{X}, \mathbf{A})}{\int p(\mathbf{Y}|\phi, \mathcal{M}, \mathbf{X}, \mathbf{A})p(\phi|\mathcal{M}, \mathbf{X}, \mathbf{A})d\phi} \\
&\stackrel{(a)}{=} \frac{\prod_{v \in \mathcal{V}} p(\mathbf{y}_v|\theta, \mathcal{M}, \mathbf{X}, \mathbf{A})p(\theta|\mathcal{M}, \mathbf{X}, \mathbf{A})}{\int \prod_{v \in \mathcal{V}} p(\mathbf{y}_v|\phi, \mathcal{M}, \mathbf{X}, \mathbf{A})p(\phi|\mathcal{M}, \mathbf{X}, \mathbf{A})d\phi} \\
&\stackrel{(b)}{=} \frac{e^{-\sum_{v \in \mathcal{V}} m_v \ell(f_\theta(\mathbf{X}, \mathbf{A}_{\mathcal{M}})_v, \mathbf{y}_v)} p(\theta|\mathcal{M}, \mathbf{X}, \mathbf{A})}{\int e^{-\sum_{v \in \mathcal{V}} m_v \ell(f_\phi(\mathbf{X}, \mathbf{A}_{\mathcal{M}})_v, \mathbf{y}_v)} p(\phi|\mathcal{M}, \mathbf{X}, \mathbf{A})d\phi}, \quad (12)
\end{aligned}$$

845 where (a) follows since, from (1), each embedding \mathbf{z}_v is a function of $(\mathbf{X}, \mathbf{A}, \theta, \mathcal{M})$; by [36,
846 Sec. 3.3], this implies that \mathbf{z}_v is independent of \mathbf{z}_u given $\mathbf{X}, \mathbf{A}, \theta, \mathcal{M}$ for all $v \neq u$. Applying
847 the softmax and indexing with \mathbf{y}_v , see (2), preserves this independence. Step (b) follows from the
848 definition of the loss function in (3). Also, the $n \times n$ matrices $\mathbf{A}_{\mathcal{M}}$ and $\mathbf{A}_{\tilde{\mathcal{M}}}$ are defined as

$$(\mathbf{A}_{\mathcal{M}})_{uw} = \begin{cases} A_{uw}, & m_u = m_w = 1, \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad (\mathbf{A}_{\tilde{\mathcal{M}}})_{uw} = \begin{cases} (\mathbf{A}_{\mathcal{M}})_{uw}, & u, w \neq v, \\ 0, & \text{otherwise} \end{cases}$$

849 for $u, w \in \mathcal{V}$. In words, $\mathbf{A}_{\mathcal{M}}$ is the adjacency matrix of the subgraph of \mathcal{G} induced by the nodes
850 included in the training set, while $\mathbf{A}_{\tilde{\mathcal{M}}}$ corresponds to the subgraph of this resulting graph obtained
851 by removing the target node and all its adjacent edges.

852 We can now evaluate the log likelihood-ratio in terms of the loss function:

$$\begin{aligned}
& \log \frac{p(\boldsymbol{\theta}|m_v = 1, \tilde{\mathcal{M}}, \mathcal{G})}{p(\boldsymbol{\theta}|m_v = 0, \tilde{\mathcal{M}}, \mathcal{G})} \\
&= - \sum_{u \in \mathcal{V}} m_u \ell(f_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{A}_{\mathcal{M}})_u, \mathbf{y}_u) + \sum_{u \in \mathcal{V} \setminus \{v\}} m_u \ell(f_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{A}_{\tilde{\mathcal{M}}})_u, \mathbf{y}_u) \\
&+ \log \frac{p(\boldsymbol{\theta}|m_v = 1, \tilde{\mathcal{M}}, \mathbf{X}, \mathbf{A})}{p(\boldsymbol{\theta}|m_v = 0, \tilde{\mathcal{M}}, \mathbf{X}, \mathbf{A})} \\
&- \log \frac{\int e^{-\sum_{u \in \mathcal{V}} m_u \ell(f_{\boldsymbol{\phi}}(\mathbf{X}, \mathbf{A}_{\mathcal{M}})_u, \mathbf{y}_u)} p(\boldsymbol{\phi}|m_v = 1, \tilde{\mathcal{M}}, \mathbf{X}, \mathbf{A}) d\boldsymbol{\phi}}{\int e^{-\sum_{u \in \mathcal{V} \setminus \{v\}} m_u \ell(f_{\boldsymbol{\phi}'}(\mathbf{X}, \mathbf{A}_{\tilde{\mathcal{M}}})_u, \mathbf{y}_u)} p(\boldsymbol{\phi}'|m_v = 0, \tilde{\mathcal{M}}, \mathbf{X}, \mathbf{A}) d\boldsymbol{\phi}'} \\
&\stackrel{(c)}{\approx} -\ell(f_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{A}_{\mathcal{M}})_v, \mathbf{y}_v) - \Delta \mathcal{L}_{\mathcal{N}_L(v)} \\
&- \log \int e^{-\ell(f_{\boldsymbol{\phi}}(\mathbf{X}, \mathbf{A}_{\mathcal{M}})_v, \mathbf{y}_v) - \Delta \mathcal{L}_{\mathcal{N}_L(v)}} \frac{e^{-\sum_{u \in \mathcal{V} \setminus \{v\}} m_u \ell(f_{\boldsymbol{\phi}}(\mathbf{X}, \mathbf{A}_{\tilde{\mathcal{M}}})_u, \mathbf{y}_u)} p(\boldsymbol{\phi}|\tilde{\mathcal{M}}, \mathbf{X}, \mathbf{A})}{\int e^{-\sum_{u \in \mathcal{V} \setminus \{v\}} m_u \ell(f_{\boldsymbol{\phi}'}(\mathbf{X}, \mathbf{A}_{\tilde{\mathcal{M}}})_u, \mathbf{y}_u)} p(\boldsymbol{\phi}'|\tilde{\mathcal{M}}, \mathbf{X}, \mathbf{A}) d\boldsymbol{\phi}'} d\boldsymbol{\phi} \\
&\stackrel{(d)}{=} -\ell(f_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{A}_{\mathcal{M}})_v, \mathbf{y}_v) - \Delta \mathcal{L}_{\mathcal{N}_L(v)} - \log \int e^{-\ell(f_{\boldsymbol{\phi}}(\mathbf{X}, \mathbf{A}_{\mathcal{M}})_v, \mathbf{y}_v) - \Delta \mathcal{L}_{\mathcal{N}_L(v)}} p(\boldsymbol{\phi}|\tilde{\mathcal{M}}, \mathcal{G}) d\boldsymbol{\phi},
\end{aligned} \tag{13}$$

853 where in step (c) we have used that $p(\boldsymbol{\theta}|m_v = 1, \tilde{\mathcal{M}}, \mathbf{X}, \mathbf{A}) \approx p(\boldsymbol{\theta}|m_v = 0, \tilde{\mathcal{M}}, \mathbf{X}, \mathbf{A}) \approx$
854 $p(\boldsymbol{\theta}|\tilde{\mathcal{M}}, \mathbf{X}, \mathbf{A})$, which follows from the asymptotic convergence properties of the model poste-
855 rior distribution given a large training set² [14, Appendix A], and we also used

$$\begin{aligned}
& - \sum_{u \in \mathcal{V}} m_u \ell(f_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{A}_{\mathcal{M}})_u, \mathbf{y}_u) + \sum_{u \in \mathcal{V} \setminus \{v\}} m_u \ell(f_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{A}_{\tilde{\mathcal{M}}})_u, \mathbf{y}_u) \\
&= -\ell(f_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{A}_{\mathcal{M}})_v, \mathbf{y}_v) - \sum_{u \in \mathcal{V} \setminus \{v\}} m_u (\ell(f_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{A}_{\mathcal{M}})_u, \mathbf{y}_u) - \ell(f_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{A}_{\tilde{\mathcal{M}}})_u, \mathbf{y}_u)) \\
&= -\ell(f_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{A}_{\mathcal{M}})_v, \mathbf{y}_v) - \sum_{u \in \mathcal{N}_L(v)} m_u (\ell(f_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{A}_{\mathcal{M}})_u, \mathbf{y}_u) - \ell(f_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{A}_{\tilde{\mathcal{M}}})_u, \mathbf{y}_u)) \\
&= -\ell(f_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{A}_{\mathcal{M}})_v, \mathbf{y}_v) - \Delta \mathcal{L}_{\mathcal{N}_L(v)}
\end{aligned}$$

856 since the inclusion or exclusion of node v affects only the predictions within its L -hop neighborhood
857 $\mathcal{N}_L(v)$, assuming an L -layer GNN. Furthermore, in step (d) we used (12) to simplify the model
858 distribution that is integrated over.

859 Combining (11) and (13) concludes the proof.

860 B.2 G-BASE: Practical Bayes-Optimal MIA against GNNs

861 We elaborate on the proposed sampling strategies used to approximate the expectation over the
862 membership statuses of all non-target samples, $\tilde{\mathcal{M}}$, i.e., the term $\mathbb{E}_{\tilde{\mathcal{M}} \sim P(\tilde{\mathcal{M}}|\boldsymbol{\theta}, \mathcal{G})}$ in Equation (5).

863 **Model-independent sampling.** The simplest approximation assumes that the membership inference
864 configuration $\tilde{\mathcal{M}}$ is independent of the target model, thereby ignoring the conditioning on $\boldsymbol{\theta}$, i.e.,
865 $P(\tilde{\mathcal{M}}|\boldsymbol{\theta}, \mathcal{G}) \approx P(\tilde{\mathcal{M}}|\mathcal{G})$. Under this assumption, we approximate $P(\tilde{\mathcal{M}}|\mathcal{G})$ by treating each member-
866 ship indicator $m_v \in \tilde{\mathcal{M}}$ as i.i.d. according to a Bernoulli distribution with parameter $\lambda = P(m_v = 1)$.
867 We then generate samples of $\tilde{\mathcal{M}}$ as $\tilde{\mathcal{M}} = \{m_v : m_v \sim \text{Ber}(\lambda), v \in \mathcal{V}\}$.

868 **MCMC sampling.** To account for the dependence of $\tilde{\mathcal{M}}$ on the target model $\boldsymbol{\theta}$, we develop a Markov
869 chain Monte Carlo (MCMC) method based on the Metropolis-Hastings algorithm. The goal is to
870 construct a Markov chain over membership configurations $\tilde{\mathcal{M}}$ with $P(\tilde{\mathcal{M}}|\boldsymbol{\theta}, \mathcal{G})$ as its stationary

²The parameter distribution of a trained model also depends on the label, in which case the result states that, in the asymptotic limit of an infinitely large dataset, there is no difference between in-models and out-models, contradicting the principles of MIAs. However, in our case, we are not conditioning on the labels, and therefore, the inclusion or exclusion of the target node gives even less information about the model parameters.

871 distribution. To apply Metropolis-Hastings, we need to be able to evaluate the unnormalized (up
 872 to a multiplicative constant) probability mass function $P(\tilde{\mathcal{M}}|\boldsymbol{\theta}, \mathcal{G})$, i.e., a function $P^*(\tilde{\mathcal{M}}|\boldsymbol{\theta}, \mathcal{G})$
 873 satisfying

$$P(\tilde{\mathcal{M}}|\boldsymbol{\theta}, \mathcal{G}) = \frac{P^*(\tilde{\mathcal{M}}|\boldsymbol{\theta}, \mathcal{G})}{\sum_{\mathcal{M}'} P^*(\mathcal{M}'|\boldsymbol{\theta}, \mathcal{G})}.$$

874 We derive such a function using Bayes' rule, assuming a uniform prior over $\tilde{\mathcal{M}}$ to eliminate the prior
 875 terms,

$$P(\tilde{\mathcal{M}}|\boldsymbol{\theta}, \mathcal{G}) = \frac{p(\boldsymbol{\theta}|\tilde{\mathcal{M}}, \mathcal{G})}{\sum_{\mathcal{M}'} p(\boldsymbol{\theta}|\mathcal{M}', \mathcal{G})}.$$

876 Using (12) for the model posterior, we obtain

$$P(\tilde{\mathcal{M}}|\boldsymbol{\theta}, \mathcal{G}) \propto \frac{e^{-\sum_{u \in \mathcal{V}} m_u \ell(f_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{A}_{\tilde{\mathcal{M}}})_u, \mathbf{y}_u)}}{\int e^{-\sum_{u \in \mathcal{V}} m_u \ell(f_{\boldsymbol{\phi}}(\mathbf{X}, \mathbf{A}_{\tilde{\mathcal{M}}})_u, \mathbf{y}_u)} p(\boldsymbol{\phi}) d\boldsymbol{\phi}} = P^*(\tilde{\mathcal{M}}|\boldsymbol{\theta}, \mathcal{G}). \quad (14)$$

877 To sample from this distribution, we initialize the Markov chain at a randomly-selected configura-
 878 tion $\tilde{\mathcal{M}}^{(0)}$, and define a step-size parameter ϵ controlling the proportion of indicator variables flipped
 879 at each iteration. At iteration t , we propose a new configuration $\tilde{\mathcal{M}}^*$ by flipping a fraction ϵ of the
 880 membership indicators in $\tilde{\mathcal{M}}^{(t)}$, and compute the log acceptance ratio

$$\begin{aligned} \log \frac{p(\boldsymbol{\theta}|\tilde{\mathcal{M}}^*, \mathcal{G})}{p(\boldsymbol{\theta}|\tilde{\mathcal{M}}^{(t)}, \mathcal{G})} &= \sum_{u \in \mathcal{V}} (m_u^{(t)} \ell(f_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{A}_{\tilde{\mathcal{M}}^{(t)}})_u, \mathbf{y}_u) - m_u^* \ell(f_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{A}_{\tilde{\mathcal{M}}^*})_u, \mathbf{y}_u)) \\ &\quad + \log \left(\int e^{-\sum_{u \in \mathcal{V}} m_u^{(t)} \ell(f_{\boldsymbol{\phi}}(\mathbf{X}, \mathbf{A}_{\tilde{\mathcal{M}}^{(t)}})_u, \mathbf{y}_u)} p(\boldsymbol{\phi}) d\boldsymbol{\phi} \right) \\ &\quad - \log \left(\int e^{-\sum_{u \in \mathcal{V}} m_u^* \ell(f_{\boldsymbol{\phi}}(\mathbf{X}, \mathbf{A}_{\tilde{\mathcal{M}}^*})_u, \mathbf{y}_u)} p(\boldsymbol{\phi}) d\boldsymbol{\phi} \right). \end{aligned} \quad (15)$$

881 Each integral can be efficiently approximated via Monte Carlo sampling using shadow models.
 882 Notably, the same shadow models used to approximate the inner expectation in (5) can be reused here
 883 to evaluate (15). We then accept the proposal with probability $\min(1, p(\boldsymbol{\theta}|\tilde{\mathcal{M}}^*, \mathcal{G})/p(\boldsymbol{\theta}|\tilde{\mathcal{M}}^{(t)}, \mathcal{G}))$.
 884 Specifically, we draw $u \sim \text{Uniform}(0, 1)$ and set

$$(\mathcal{M}^{(t+1)}, p(\boldsymbol{\theta}|\tilde{\mathcal{M}}^{(t+1)}, \mathcal{G})) = \begin{cases} (\mathcal{M}^*, p(\boldsymbol{\theta}|\tilde{\mathcal{M}}^*, \mathcal{G})) & \text{if } \frac{p^*}{p^{(t)}} > u \\ (\mathcal{M}^{(t)}, p(\boldsymbol{\theta}|\mathcal{M}^{(t)}, \mathcal{G})) & \text{otherwise.} \end{cases}$$

885 To obtain approximately independent samples, we insert a burn-in period at the beginning of the
 886 chain and use thinning—collecting samples only after a sufficient number of iterations.

887 **0-hop MIA sampling.** Recall that $P(\tilde{\mathcal{M}}|\boldsymbol{\theta}, \mathcal{G})$ is the joint distribution of the membership of all non-
 888 target nodes. A natural approximation is to apply a per-node MIA to estimate individual membership
 889 probabilities and assume independence across nodes. Concretely, we apply a 0-hop MIA—i.e., we
 890 ignore the graph structure and treat nodes as i.i.d. samples. This reduces the problem to the standard
 891 i.i.d. setting, where well-established MIA methods can be applied. Any i.i.d.-based MIA that yields
 892 membership scores convertible to probabilities can be used. Here, we adopt the attack introduced in
 893 Section 5.2 which approximates the Bayes-optimal inference rule in the i.i.d setting. We then generate
 894 samples from the approximate posterior as $\tilde{\mathcal{M}} = \{m_v : m_v \sim \text{Ber}(P(m_v|\boldsymbol{\theta}, \mathbf{X}_v, \mathbf{y}_v)), v \in \mathcal{V}\}$,
 895 where $P(m_v|\boldsymbol{\theta}, \mathbf{X}_v, \mathbf{y}_v)$ is the membership probability of node v assigned by the BASE attack.

896 C Selecting the Decision Threshold

897 Since an adversary does not have access to ground truth membership labels, they cannot directly tune
 898 the decision threshold (by sweeping τ) to achieve a specific FPR. Instead, the adversary must choose
 899 a threshold that is expected to yield an FPR close to the maximal tolerated FPR. For MIAs based on
 900 shadow models, we propose finding such a threshold by designating a subset of the shadow models
 901 as simulated target models. Because their ground truth membership statuses are known, they can
 902 be attacked using the remaining shadow models. This allows the adversary to sweep over decision
 903 thresholds and identify the values that is expected to approximately yield the desired FPR.

Table 3: Decision thresholds for RMIA, BASE and G-BASE resulting in the largest FPR less than or equal to the target FPR (1% and 0.1%). The threshold is reported as the mean \pm standard deviation, averaged over 10 different target models and sets of target samples. A comparatively small standard deviation indicates that the threshold is not expected to vary too much over different target models, allowing the adversary to estimate a threshold using shadow models as simulated target models. The RMIA threshold does not satisfy the empirical rule “threshold = 1 - FPR” as reported in [13].

| K | ATTACK | CORA (GCN) | | CITSEER (GAT) | | PUBMED (GRAPHSAGE) | |
|-----|--------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| | | THRESHOLD@FPR | | THRESHOLD@FPR | | TPR@FPR | |
| | | 1% | 0.1% | 1% | 0.1% | 1% | 0.1% |
| 8 | RMIA | 0.8566 \pm 0.0078 | 0.9030 \pm 0.0174 | 0.8240 \pm 0.0124 | 0.8990 \pm 0.0299 | 0.9474 \pm 0.0039 | 0.9822 \pm 0.0030 |
| | BASE | 0.6024 \pm 0.0098 | 0.6913 \pm 0.0342 | 0.5907 \pm 0.0097 | 0.6971 \pm 0.0548 | 0.6052 \pm 0.0093 | 0.7140 \pm 0.0161 |
| | G-BASE (MIA) | 0.5619 \pm 0.0072 | 0.6204 \pm 0.0327 | 0.5724 \pm 0.0059 | 0.6471 \pm 0.0309 | 0.5928 \pm 0.0053 | 0.6771 \pm 0.0175 |
| 4 | RMIA (OFF) | 0.8275 \pm 0.0087 | 0.8612 \pm 0.0128 | 0.7649 \pm 0.0101 | 0.8315 \pm 0.0145 | 0.9719 \pm 0.0027 | 0.9926 \pm 0.0016 |
| | BASE (OFF) | 0.5358 \pm 0.0062 | 0.6015 \pm 0.0454 | 0.5540 \pm 0.0075 | 0.6341 \pm 0.0368 | 0.6240 \pm 0.0093 | 0.7422 \pm 0.0182 |
| | G-BASE (OFF) | 0.7034 \pm 0.0127 | 0.8160 \pm 0.0324 | 0.7064 \pm 0.0168 | 0.8466 \pm 0.0311 | 0.6661 \pm 0.0058 | 0.8031 \pm 0.0213 |

| K | ATTACK | FLICKR (GCN) | | AMAZON-PHOTO (GAT) | | GITHUB (GRAPHSAGE) | |
|-----|--------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| | | THRESHOLD@FPR | | THRESHOLD@FPR | | THRESHOLD@FPR | |
| | | 1% | 0.1% | 1% | 0.1% | 1% | 0.1% |
| 8 | RMIA | 0.8029 \pm 0.0086 | 0.9142 \pm 0.0099 | 0.9267 \pm 0.0101 | 0.9675 \pm 0.0144 | 0.9462 \pm 0.0053 | 0.9785 \pm 0.0038 |
| | BASE | 0.9080 \pm 0.0091 | 0.9873 \pm 0.0036 | 0.7167 \pm 0.0392 | 0.8866 \pm 0.0762 | 0.6328 \pm 0.0178 | 0.7250 \pm 0.0300 |
| | G-BASE (MIA) | 0.6770 \pm 0.0108 | 0.7910 \pm 0.0109 | 0.6031 \pm 0.0188 | 0.7316 \pm 0.0532 | 0.6162 \pm 0.0114 | 0.6990 \pm 0.0189 |
| 4 | RMIA (OFF) | 0.8695 \pm 0.0074 | 0.9330 \pm 0.0042 | 0.8679 \pm 0.0084 | 0.9032 \pm 0.0055 | 0.9490 \pm 0.0119 | 0.9801 \pm 0.0056 |
| | BASE (OFF) | 0.9143 \pm 0.0084 | 0.9884 \pm 0.0036 | 0.5806 \pm 0.0119 | 0.7219 \pm 0.0830 | 0.6317 \pm 0.0233 | 0.7421 \pm 0.0258 |
| | G-BASE (OFF) | 0.6965 \pm 0.0109 | 0.8072 \pm 0.0191 | 0.6822 \pm 0.0121 | 0.8604 \pm 0.0279 | 0.6677 \pm 0.0108 | 0.7826 \pm 0.0184 |

By repeating the threshold estimation process across multiple simulated target models, the adversary can assess the variability in the resulting thresholds. A conservative adversary would choose a threshold at least as large as the maximum threshold obtained. Using the mean of the thresholds computed against the simulated target models is another viable option. As shown in Table 3 (graph data) and Table 4 (i.i.d. data), the variation in threshold values across target models is small, indicating that thresholds estimated from simulated target models are fairly stable. RMIA exhibits the lowest threshold variance. The viability of estimating the threshold using simulated target models also depends on how sensitive the resulting FPR is to threshold fluctuations. To demonstrate the effectiveness of this approach, we train 10 shadow models to act as simulated target models. These simulated target models are attacked and, using the ground truth knowledge about the training membership status, the decision thresholds resulting in an FPR not exceeding 1% are computed. The estimated threshold is then taken as the average of these decision thresholds. To run the attacks, 8 (online) and 4 (offline) separate shadow models are used. Table 5 shows the TPR and FPR achieved when using the estimated threshold on real target models. The FPR obtained using the estimated threshold does not deviate much from the target 1% FPR. Hence, the TPR is also close to the TPR of the exact 1% FPR threshold. Despite the differences in threshold variance between LIRA, RMIA, and our attacks (see Table 3 and Table 4), there is no significant difference in the accuracy of the estimated threshold.

We conclude with a remark on alternative methods for threshold selection. For LIRA, the threshold can be selected from the fitted Gaussian distribution. However, the accuracy of the estimated threshold depends on how well the Gaussian distribution fits the logit-scaled confidence values. As such, this approach also relies on an estimation based on population data, and on top, the heuristic observation that logit-scaled confidence values often look normal distributed. For RMIA with $\gamma = 1$, the authors argue that their attack is calibrated such that a threshold $\beta = 1 - \alpha$ results in FPR α . Investigating this heuristic rule further, we find that it generally does not hold. In Table 3 and Table 4 we see that the thresholds resulting in an FPR at most 1% or 0.1% are lower than $\beta = 0.99$ or $\beta = 0.999$, respectively, across all datasets. Moreover, Table 6 (graph data) and Table 7 (i.i.d data) show that the actual FPR obtained when setting $\beta = 0.9$ and $\beta = 0.99$ is lower than 10% and 1%, respectively, as would be expected if $\beta = 1 - \alpha$ where to give FPR α . Consequently, the TPR obtained at $\beta = 0.9$ and $\beta = 0.99$ is significantly lower than what is possible to achieve at FPR 10% and 1%, respectively. As an example, at FPR 1%, online RMIA achieves a mean TPR of 24.00% on Citeseer (see Table 1), whereas using the threshold $\beta = 0.99$ instead results in no true positives at all.

D Membership Inference Game for i.i.d. Data

The following definition of a membership inference game closely follows the one in [11, Def. 1] and the ones in [14].

Table 4: Decision thresholds for BASE, RMIA and LiRA resulting in the largest FPR less than or equal to the target FPR (1% and 0.1%). The threshold is reported as the mean \pm standard deviation, averaged over 10 different target models and sets of target samples. A comparatively small standard deviation indicates that the threshold is not expected to vary too much over different target models, allowing the adversary to estimate a threshold using shadow models as simulated target models. The RMIA threshold does not satisfy the empirical rule threshold=1-FPR as reported in [13].

| K | ATTACK | CIFAR-10 | | CIFAR-100 | |
|----|------------|----------------------|----------------------|----------------------|----------------------|
| | | THRESHOLD@FPR | | THRESHOLD@FPR | |
| | | 1% | 0.1% | 1% | 0.1% |
| 32 | BASE | 0.5385 \pm 0.0473 | 0.8159 \pm 0.0669 | 0.8417 \pm 0.0735 | 1.2240 \pm 0.0914 |
| | RMIA | 0.9624 \pm 0.0047 | 0.9900 \pm 0.0028 | 0.9429 \pm 0.0139 | 0.9833 \pm 0.0061 |
| | LiRA | 0.6505 \pm 0.0515 | 1.1098 \pm 0.1126 | 0.8366 \pm 0.0478 | 1.4039 \pm 0.0804 |
| 16 | BASE (OFF) | 0.1766 \pm 0.0279 | 0.3485 \pm 0.0472 | 0.1920 \pm 0.0440 | 0.4066 \pm 0.0525 |
| | RMIA (OFF) | 0.9645 \pm 0.0050 | 0.9909 \pm 0.0023 | 0.9647 \pm 0.0078 | 0.9918 \pm 0.0024 |
| | LiRA (OFF) | -0.0837 \pm 0.0223 | -0.0323 \pm 0.0136 | -0.0770 \pm 0.0201 | -0.0210 \pm 0.0080 |
| 8 | BASE | 0.6116 \pm 0.0601 | 0.9740 \pm 0.0839 | 0.9207 \pm 0.0864 | 1.4146 \pm 0.1155 |
| | RMIA | 0.9675 \pm 0.0035 | 0.99322 \pm 0.0015 | 0.9505 \pm 0.0102 | 0.9884 \pm 0.0037 |
| | LiRA | 0.9362 \pm 0.0812 | 1.5689 \pm 0.1635 | 1.1176 \pm 0.0898 | 2.0389 \pm 0.1434 |
| 4 | BASE (OFF) | 0.2312 \pm 0.0364 | 0.4795 \pm 0.0571 | 0.24979 \pm 0.0569 | 0.5437 \pm 0.0737 |
| | RMIA (OFF) | 0.9698 \pm 0.0032 | 0.9934 \pm 0.0014 | 0.9689 \pm 0.0059 | 0.9942 \pm 0.0021 |
| | LiRA (OFF) | -0.0664 \pm 0.0199 | -0.0221 \pm 0.0101 | -0.0581 \pm 0.0162 | -0.0132 \pm 0.0056 |

Table 5: Attack performance using a threshold estimated by attacking 10 simulated target models. The estimated threshold is the average 1% FPR threshold against the simulated target models. TPR at 1% FPR is reported for comparison. Performance is measured as mean \pm standard deviation against 10 target models. The FPR against the target models when using the estimated threshold is close to 1%. Consequently, the TPR at the estimated threshold is close to the TPR at the 1% FPR threshold. This method of estimating the threshold at a given fixed FPR does work for our attacks, LiRA, and RMIA.

| K | ATTACK | CORA (GCN) | | | CITSEER (GAT) | | | PUBMED (GRAPHSAGE) | | |
|---|--------------|------------------------------------|------------------------------------|-----------------|------------------------------------|------------------------------------|-----------------|-----------------------------------|-----------------------------------|-----------------|
| | | TPR@1%FPR (%) | ESTIMATED THRESHOLD | | TPR@1%FPR (%) | ESTIMATED THRESHOLD | | TPR@1%FPR (%) | ESTIMATED THRESHOLD | |
| | | | TPR (%) | FPR (%) | | TPR (%) | FPR (%) | | TPR (%) | FPR (%) |
| 8 | LiRA | 9.25 \pm 3.79 | 6.97 \pm 0.65 | 0.83 \pm 0.34 | 15.35 \pm 3.20 | 12.78 \pm 0.98 | 0.78 \pm 0.30 | 1.22 \pm 0.24 | 1.24 \pm 0.14 | 1.04 \pm 0.13 |
| | RMIA | 15.83 \pm 2.57 | 16.26 \pm 2.00 | 1.05 \pm 0.28 | 24.08 \pm 2.03 | 21.46 \pm 2.36 | 0.85 \pm 0.22 | 3.07 \pm 0.31 | 2.77 \pm 0.15 | 0.87 \pm 0.11 |
| | BASE | 15.88 \pm 2.57 | 17.19 \pm 1.29 | 1.08 \pm 0.39 | 24.09 \pm 2.03 | 22.24 \pm 1.13 | 0.87 \pm 0.10 | 3.07 \pm 0.31 | 2.87 \pm 0.22 | 0.90 \pm 0.13 |
| | G-BASE | 14.00 \pm 3.11 | 14.96 \pm 1.70 | 1.03 \pm 0.30 | 21.03 \pm 2.90 | 21.85 \pm 1.50 | 1.14 \pm 0.30 | 5.33 \pm 0.58 | 5.39 \pm 0.39 | 1.01 \pm 0.12 |
| 4 | LiRA (OFF) | 12.45 \pm 3.60 | 11.73 \pm 1.30 | 0.84 \pm 0.30 | 18.17 \pm 2.24 | 20.26 \pm 0.86 | 1.22 \pm 0.30 | 1.52 \pm 0.19 | 2.04 \pm 0.39 | 1.37 \pm 0.29 |
| | RMIA (OFF) | 16.94 \pm 3.09 | 17.39 \pm 1.14 | 1.09 \pm 0.36 | 25.38 \pm 3.42 | 23.87 \pm 0.91 | 0.89 \pm 0.32 | 3.41 \pm 0.47 | 3.14 \pm 0.27 | 0.88 \pm 0.15 |
| | BASE (OFF) | 17.36 \pm 3.34 | 17.58 \pm 1.14 | 1.03 \pm 0.30 | 26.02 \pm 3.38 | 24.57 \pm 1.01 | 0.97 \pm 0.22 | 3.30 \pm 0.39 | 3.17 \pm 0.24 | 0.91 \pm 0.18 |
| | G-BASE (OFF) | 11.31 \pm 2.97 | 11.88 \pm 1.18 | 1.15 \pm 0.45 | 16.82 \pm 2.72 | 15.75 \pm 0.80 | 0.97 \pm 0.31 | 5.60 \pm 0.40 | 5.55 \pm 0.31 | 1.00 \pm 0.11 |

Table 6: TPR and FPR at fixed threshold β for the RMIA attack with $\gamma = 1$, using the full population as the Z set. Setting $\beta = 1 - \alpha$ does result in a significantly lower FPR than α , at the cost of a lower TPR than what is possible to achieve at FPR α . 256 shadow models are used for the online attack, and 128 for the offline attack.

| ATTACK | CORA (GCN) | | | | CITSEER (GAT) | | | |
|------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| | $\beta = 0.9$ | | $\beta = 0.99$ | | $\beta = 0.9$ | | $\beta = 0.99$ | |
| | TPR (%) | FPR (%) | TPR (%) | FPR (%) | TPR (%) | FPR (%) | TPR (%) | FPR (%) |
| RMIA | 2.73 \pm 0.49 | 0.00 \pm 0.00 | 0.00 \pm 0.00 | 0.00 \pm 0.00 | 3.31 \pm 0.71 | 0.01 \pm 0.04 | 0.00 \pm 0.00 | 0.00 \pm 0.00 |
| RMIA (OFF) | 0.00 \pm 0.00 | 0.00 \pm 0.00 | 0.00 \pm 0.00 | 0.00 \pm 0.00 | 0.10 \pm 0.09 | 0.00 \pm 0.00 | 0.00 \pm 0.00 | 0.00 \pm 0.00 |

939 **Definition 5. (Membership inference game)**

- 940 1. The challenger samples a dataset $\mathcal{D}_{train} \subset \mathcal{D}$ from a data population pool \mathcal{D} and trains a
941 model θ on \mathcal{D}_{train} .
- 942 2. The challenger flips a fair coin to generate a bit b . If $b = 0$, a data point v is randomly selected
943 from $\mathcal{D} \setminus \mathcal{D}_{train}$. If $b = 1$, the data point v is selected from \mathcal{D}_{train} .
- 944 3. The challenger gives the adversary the population pool \mathcal{D} , the target sample v , and black-box
945 access to the trained model f_θ .
- 946 4. The adversary may also have access to additional side information (such as knowledge about the
947 training algorithm or model architecture). Using this information, the adversary performs a MIA
948 $\hat{m}_v \leftarrow MIA(f_\theta, v, \mathcal{D}, \mathcal{H})$, where \hat{m}_v is an estimate of the membership status of sample v , m_v .
- 949 5. The attack is successful on a data point v if $\hat{m}_v = m_v$.

Table 7: TPR and FPR values for RMIA and RMIA (OFF) on CIFAR-10 at decision thresholds $\beta = 1 - \alpha$ where $\alpha \in \{0.01, 0.001\}$. Results are reported as mean \pm standard deviation over 10 runs.

| K | ATTACK | $\beta = 0.99$ | | $\beta = 0.999$ | |
|-----|------------|-----------------|-----------------|-----------------|-----------------|
| | | TPR (%) | FPR (%) | TPR (%) | FPR (%) |
| 32 | RMIA | 1.90 ± 0.09 | 0.11 ± 0.04 | 0.20 ± 0.03 | 0.00 ± 0.00 |
| 16 | RMIA (OFF) | 1.89 ± 0.11 | 0.13 ± 0.06 | 0.21 ± 0.03 | 0.00 ± 0.01 |
| 8 | RMIA | 1.82 ± 0.07 | 0.18 ± 0.04 | 0.20 ± 0.04 | 0.01 ± 0.00 |
| 4 | RMIA (OFF) | 1.82 ± 0.09 | 0.20 ± 0.05 | 0.20 ± 0.02 | 0.01 ± 0.01 |

E BASE: Practical Bayes-Optimal MIA for i.i.d. Data

E.1 Proof of Corollary 1

The i.i.d. setting corresponds to $\mathbf{A} = \mathbf{I}$, where \mathbf{I} is the identity matrix. In this case, since the data points are independent, the neighborhood-dependent term (7) vanishes, $\mathcal{G} = \mathcal{D}$, and the membership indicator satisfies $m_v = 1$ if the data sample $v \in [n]$ is included in the training set and $m_v = 0$ otherwise. Hence, $S_{\mathcal{L}}(\boldsymbol{\theta}, \mathcal{D}, m_v)$ in (6) reduces to the individual loss value of the target sample v , i.e., $S_{\mathcal{L}}(\boldsymbol{\theta}, \mathcal{D}, m_v) = \ell(f_{\boldsymbol{\theta}}(x_v), \mathbf{y}_v)$.

E.2 BASE: Practical Bayes-Optimal MIA for i.i.d. Data

Recall that under the i.i.d. assumption, we have $\mathcal{G} = \mathcal{D}$ and the loss-based signal simplifies to $S_{\mathcal{L}}(\boldsymbol{\theta}, \mathcal{D}, m_v) = \ell(f_{\boldsymbol{\theta}}(x_v), \mathbf{y}_v)$ in Theorem 1. Now, since the loss-based signal no longer depends on $\tilde{\mathcal{M}}$, the only dependence on $\tilde{\mathcal{M}}$ in Bayes-optimal membership inference rule is contained in the posterior model distribution $p(\phi | \tilde{\mathcal{M}}, \mathcal{D})$. By approximating this posterior model distribution by the prior $p(\phi)$, we have removed all the dependence on $\tilde{\mathcal{M}}$ in the approximation and the outer expectation over $\tilde{\mathcal{M}}$ is trivial.

The posterior membership probability $P(m_v | \boldsymbol{\theta}, \mathcal{D})$ then simplifies to

$$P(m_v | \boldsymbol{\theta}, \mathcal{D}) = \sigma \left(-\ell(f_{\boldsymbol{\theta}}(x_v), \mathbf{y}_v) - \log \int e^{-\ell(f_{\phi}(x_v), \mathbf{y}_v)} p(\phi) d\phi + \log \frac{\lambda}{1 - \lambda} \right). \quad (16)$$

The remaining expectation over the prior model distribution is still intractable but can be efficiently approximated using Monte Carlo sampling of shadow models,

$$\log \int e^{-\ell(f_{\phi}(x_v), \mathbf{y}_v)} p(\phi) d\phi \approx \log \left(\frac{1}{K} \sum_{k=1}^K e^{-\ell(f_{\phi_k}(x_v), \mathbf{y}_v)} \right), \quad \phi_k \leftarrow \mathcal{T}(\mathcal{D}_k), \quad (17)$$

where the shadow models are trained on sampled datasets \mathcal{D}_k from the data population, in analogy with (8). Substituting the Monte Carlo approximation from (17) into (16), we can formalize the resulting attack as in Definition 3 (Section 5.2).

Note that neither the sigmoid function nor the membership prior term are necessary for the attack. As shown in Lemma 1, an equivalent attack can be obtained by applying the inverse sigmoid function and subtracting the prior term, which corresponds to a monotonic transformation. However, we retain both the sigmoid function and prior term to preserve the interpretation of the attack score as a posterior probability.

E.3 Proof of Theorem 2

We begin by establishing the following result.

Lemma 1. *Two score-based MIAs are equivalent if their score functions are related by a monotonic transformation.*

Proof. Let \mathbf{a} and \mathbf{b} denote the vector of prediction scores for two MIAs A and B , respectively, after attacking an arbitrary target model using N arbitrary target samples. Furthermore, let τ_A be an arbitrary decision threshold for attack A , such that the positive predictions are $\mathcal{M}_A = \{i :$

982 $i \in \{1, \dots, N\}, \mathbf{a}_i > \tau_A\}$. By assumption, there exists a strictly increasing function g such that
 983 $g(\mathbf{a}_i) = \mathbf{b}_i$ for all $i \in \{1, \dots, N\}$. Now let $\tau_B = g(\tau_A)$, then the positive predictions of attack B
 984 are given by

$$\begin{aligned}\mathcal{M}_B &= \{i : i \in \{1, N\}, \mathbf{b}_i > \tau_B\} \\ &= \{i : i \in \{1, N\}, g(\mathbf{a}_i) > g(\tau_A)\} \\ &= \{i : i \in \{1, N\}, \mathbf{a}_i > \tau_A\} \\ &= \mathcal{M}_A,\end{aligned}$$

985 where the third equality follows from the fact that $g(x) > g(y)$ if and only if $x > y$ for a strictly
 986 increasing function g . Since τ_A was arbitrary, A and B are equivalent by Definition 4. \square

987 The intuition behind the result of Lemma 1 is that a monotonic transformation preserves the order of
 988 the membership scores. When a decision threshold is applied to MIA scores, only the target samples
 989 with top- k highest score (with k depending on the threshold) are classified as members. However,
 990 since the order of the scores is preserved, the top- k scores will correspond to the same target samples
 991 for both MIAs.

992 Equipped with Lemma 1, we are now ready to prove Theorem 2. In particular, since the composition
 993 of monotonic transformations is itself a monotonic transformation, it follows that Lemma 1 also
 994 applies when the score functions are related by such a composition, which we will use repeatedly in
 995 the proof that follows.

996 The RMIA score function is defined as

$$\Lambda_{\text{RMIA}}(\mathbf{x}_i, y_i; \boldsymbol{\theta}) = \Pr_{(\mathbf{x}_j, y_j) \sim \pi} \left[\frac{p(\boldsymbol{\theta}|\mathbf{x}_i, y_i)}{p(\boldsymbol{\theta}|\mathbf{x}_j, y_j)} \geq \gamma \right], \quad (18)$$

997 where $\boldsymbol{\theta}$ is the target model parameters, (\mathbf{x}_i, y_i) the feature-label pair defining the target sample, and
 998 π the data population.

999 To prove that BASE is equivalent to RMIA when $\gamma = 1$, it suffices (by Lemma 1) to show that the
 1000 two attacks produce prediction scores that are related by a monotonic transformation. We do this in
 1001 two steps:

- 1002 1. We show that BASE is equivalent to an attack that computes the ratio between the target
 1003 model's confidence value and the expected confidence value over a distribution of shadow
 1004 models. We refer to this attack as the *mean confidence attack* (MCA).
- 1005 2. We derive the monotonic transformation that relates MCA to RMIA, which turns out to
 1006 be a cumulative distribution function (CDF), restricted to a domain determined by the data
 1007 population.

1008 The two steps are detailed in the following.

1009 **1.** The BASE score function is given by

$$\Lambda_{\text{BASE}}(\mathbf{x}_i, y_i; \boldsymbol{\theta}) = \sigma \left(-\ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) - \log \int e^{-\ell(f_{\boldsymbol{\phi}}(\mathbf{x}_i), y_i)} p(\boldsymbol{\phi}) d\boldsymbol{\phi} + \log \frac{\lambda}{1 - \lambda} \right). \quad (19)$$

1010 Since the sigmoid function is strictly increasing, it can be removed to obtain an equivalent attack. After
 1011 removing the sigmoid, the prior term becomes an additive constant, which also defines a monotonic
 1012 transformation and can therefore be discarded. Applying the (strictly increasing) exponential function
 1013 to the resulting score function, we obtain

$$\Lambda_{\text{MCA}}(\mathbf{x}_i, y_i; \boldsymbol{\theta}) = \frac{f_{\boldsymbol{\theta}}(\mathbf{x}_i)_{y_i}}{\int f_{\boldsymbol{\phi}}(\mathbf{x}_i)_{y_i} p(\boldsymbol{\phi}) d\boldsymbol{\phi}} = \frac{f_{\boldsymbol{\theta}}(\mathbf{x}_i)_{y_i}}{\mathbb{E}_{\boldsymbol{\phi}}[f_{\boldsymbol{\phi}}(\mathbf{x}_i)_{y_i}]} \quad (20)$$

1014 where we have used that the confidence is related to the negative log-likelihood loss function by
 1015 $f_{\boldsymbol{\theta}}(\mathbf{x}_i)_{y_i} = e^{-\mathcal{L}(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i)}$.

1016 **2.** Next, we show that the MCA score in (20) is related to the RMIA score function in (18) when
 1017 $\gamma = 1$. Applying Bayes' rule to the likelihood ratio in RMIA, we can rewrite the score function as

$$\begin{aligned}
 \Lambda_{\text{RMIA}}(\mathbf{x}_i, y_i; \boldsymbol{\theta}) &= \Pr_{(\mathbf{x}_j, y_j) \sim \pi} \left[\frac{p(y_i | \mathbf{x}_i, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(y_j | \mathbf{x}_j, \boldsymbol{\theta}) p(\boldsymbol{\theta})} \frac{p(y_j | \mathbf{x}_j)}{p(y_i | \mathbf{x}_i)} \geq 1 \right] \\
 &= \Pr_{(\mathbf{x}_j, y_j) \sim \pi} \left[\frac{p(y_i | \mathbf{x}_i, \boldsymbol{\theta})}{p(y_i | \mathbf{x}_i)} \geq \frac{p(y_j | \mathbf{x}_j, \boldsymbol{\theta})}{p(y_j | \mathbf{x}_j)} \right] \\
 &= \Pr_{(\mathbf{x}_j, y_j) \sim \pi} \left[\frac{p(y_i | \mathbf{x}_i, \boldsymbol{\theta})}{\mathbb{E}_{\phi}[p(y_i | \mathbf{x}_i, \phi)]} \geq \frac{p(y_j | \mathbf{x}_j, \boldsymbol{\theta})}{\mathbb{E}_{\phi}[p(y_j | \mathbf{x}_j, \phi)]} \right] \\
 &= \Pr_{(\mathbf{x}_j, y_j) \sim \pi} \left[\frac{f_{\boldsymbol{\theta}}(\mathbf{x}_i)_{y_i}}{\mathbb{E}_{\phi}[f_{\phi}(\mathbf{x}_i)_{y_i}]} \geq \frac{f_{\boldsymbol{\theta}}(\mathbf{x}_j)_{y_j}}{\mathbb{E}_{\phi}[f_{\phi}(\mathbf{x}_j)_{y_j}]} \right]. \tag{21}
 \end{aligned}$$

1018 This expression corresponds to the CDF of the random variable $\mathbf{X} = \frac{f_{\boldsymbol{\theta}}(\mathbf{x}_j)_{y_j}}{\mathbb{E}_{\phi}[f_{\phi}(\mathbf{x}_j)_{y_j}]}$ evaluated at
 1019 $\Lambda_{\text{MCA}}(\mathbf{x}_i, y_i; \boldsymbol{\theta}) = \frac{f_{\boldsymbol{\theta}}(\mathbf{x}_i)_{y_i}}{\mathbb{E}_{\phi}[f_{\phi}(\mathbf{x}_i)_{y_i}]}$. While a CDF is always non-decreasing, it is not necessarily strictly
 1020 increasing. However, since the target sample (\mathbf{x}_i, y_i) is also part of the data population pool π ,
 1021 (21) is strictly increasing on the relevant domain. Specifically, as a function of the real variable
 1022 $\frac{f_{\boldsymbol{\theta}}(\mathbf{x}_i)_{y_i}}{\mathbb{E}_{\phi}[f_{\phi}(\mathbf{x}_i)_{y_i}]}$, the CDF (21) can only be constant on intervals where there is no probability mass or
 1023 density. Therefore, RMIA and MCA are equivalent by Lemma 1.

1024 Combining steps 1 and 2 proves that BASE is equivalent to RMIA with $\gamma = 1$.

1025 **F Online vs. Offline Attacks**

1026 MIAs that use shadow models to estimate distributions of models can be designed as online or offline
 1027 attacks. In the online setting, the adversary can train shadow models on the target sample. This
 1028 requires no additional assumptions, as the adversary controls shadow model training and can always
 1029 include the target. However, in certain auditing scenarios, the online setting may be impractical—it
 1030 would necessitate retraining shadow models for every new audit target. In contrast, the offline setting
 1031 assumes that shadow models are trained once, without using any target samples. This setting can
 1032 be of importance when all the target samples are not determined when setting up the attack. Due to
 1033 the shadow model training trick introduced in [11] (see also Appendix A), it is possible to perform
 1034 efficient privacy audits in the online setting, albeit at a higher cost.

1035 The BASE attack is ideally performed in online mode, where half of the shadow models can be
 1036 trained on the target sample. Following common terminology, we refer to the shadow models trained
 1037 on the target sample as *in-models* and otherwise they are referred to as *out-models*. To compensate
 1038 for the lack of in-models in the offline setting, we introduce a scaling factor $\alpha \in [0, 1]$ on the Monte
 1039 Carlo loss term over shadow models,

$$P(m_v | \boldsymbol{\theta}, \mathcal{D}) = \sigma \left(-\ell(f_{\boldsymbol{\theta}}(\mathbf{x}_v), \mathbf{y}_v) - \alpha \log \left(\frac{1}{K} \sum_{k=1}^K e^{-\ell(f_{\phi_k}(\mathbf{x}_v), \mathbf{y}_v)} \right) + \log \frac{\lambda}{1 - \lambda} \right). \tag{22}$$

1040 Intuitively, the loss value on the target sample is expected to be smaller when using an in-model
 1041 compared to an out-model. Thus, the LogSumExp over shadow models is expected to be a negative
 1042 number of greater magnitude when computed only over out-models in offline mode, as compared
 1043 to the mix of in- and out-models used in online mode. The scaling factor reduces the magnitude of
 1044 this term to better match the expected value when using in-models. Note that this assumes that the
 1045 relative scale between the terms in the online and offline case is independent of the target sample.

1046 The same offline hyperparameter α can also be used for the G-BASE attack (see Definition 2).
 1047 However, the loss signal in the graph case also depends on the local neighborhood, which in general
 1048 contains mixed membership statuses, different for each shadow model. Hence, a shadow model
 1049 trained on the target node, but trained on just a few of the other nodes in the local neighborhood, may
 1050 attain a relatively high loss signal. On the other hand, a shadow model not trained on the target node,
 1051 but trained on a large fraction of the local neighborhood, could yield a lower loss signal. Due to this

1052 complication, scaling the integral term over shadow models is not expected to work as well as in the
 1053 0-hop or i.i.d. case. We found that a larger $\alpha \in [0.9, 1]$, meaning a smaller rescaling, works well on
 1054 the datasets considered in this work.

1055 G Differential Privacy Bound for Bayes-Optimal Membership Inference

1056 It is straightforward to bound the Bayes-optimal membership inference probability in terms of ϵ -
 1057 differential privacy (DP) [1]. DP is a mathematical framework that defines a notion of privacy for
 1058 individual data records through a measure of indistinguishability. It was originally proposed in the
 1059 context of databases, formalizing the intuitive notion that a query function on a database is private if
 1060 the inclusion or exclusion of a single data record only affects the query output by a small amount.

1061 **Definition 6. (ϵ -DP)** A randomized mechanism \mathcal{M} satisfies ϵ -DP if for any two datasets D and D'
 1062 differing in a single data sample, and any event $E \subset \text{Range}(\mathcal{M})$, it holds that

$$\log \frac{P(\mathcal{M}(D) \in E)}{P(\mathcal{M}(D') \in E)} \leq \epsilon. \quad (23)$$

1063 DP is often used as a guarantee for private machine learning. Consider a ϵ -DP training algorithm \mathcal{T}
 1064 that outputs a set of weights θ given a training dataset \mathcal{D} : $\theta \leftarrow \mathcal{T}(\mathcal{D})$. Given \mathcal{D} , the inclusion and
 1065 exclusion of the target sample v result in two datasets that differ only in one sample. Therefore,
 1066 the condition for ϵ -DP directly results in the following bound on the Bayes-optimal membership
 1067 probability (11):

$$P(m_v = 1 | \theta, \mathcal{D}) = \mathbb{E}_{\tilde{\mathcal{M}} \sim P(\tilde{\mathcal{M}} | \theta, \mathcal{D})} \left[\sigma \left(\log \frac{p(\theta | m_v = 1, \tilde{\mathcal{M}}, \mathcal{D})}{p(\theta | m_v = 0, \tilde{\mathcal{M}}, \mathcal{D})} + \log \frac{\lambda}{1 - \lambda} \right) \right] \quad (24)$$

$$\leq \mathbb{E}_{\tilde{\mathcal{M}} \sim P(\tilde{\mathcal{M}} | \theta, \mathcal{D})} \left[\sigma \left(\epsilon + \log \frac{\lambda}{1 - \lambda} \right) \right] \quad (25)$$

$$= \sigma \left(\epsilon + \log \frac{\lambda}{1 - \lambda} \right). \quad (26)$$

1068 As $\epsilon \rightarrow 0$, the upper bound approaches λ . That is, the stronger the DP guarantee, the closer the
 1069 membership inference to a random guess using only the prior. This ϵ -DP bound is particularly simple
 1070 when $\lambda = 0.5$, i.e., prior to observing the model, we are maximally uncertain about the membership
 1071 status of the target sample. In this case, the bound becomes $P(m_v = 1 | \theta, \mathcal{D}) \leq \sigma(\epsilon)$.

1072 H Additional Experiments

1073 In this appendix, we present additional experiments and results. In Section H.1, we include ROC
 1074 curves in the low FPR regime for some other dataset-target model combinations not presented in
 1075 Table 1. Further results for the mismatched adversary setting are presented in Section H.2. We
 1076 compare our different sampling strategies for G-BASE in Section H.3. Finally, in Section H.4,
 1077 we present further results on the CIFAR-10 and CIFAR-100 datasets, which are widely adopted
 1078 benchmarks for MIAs on i.i.d. data.

1079 H.1 ROC curves

1080 To compare the attack performances in terms of TPR over all low FPRs, we provide average ROC
 1081 curves in Figures 4 to 6. We see that in the online setting, the ROC curves of BASE and RMIA are
 1082 identical, as is expected in light of Theorem 2 (we use $\gamma = 1$ in RMIA). In the offline setting, BASE
 1083 and RMIA perform similarly, but not equivalently. G-BASE achieves the best performance in terms
 1084 of TPR at low FPR in all cases except Pubmed with a GCN target model in the offline setting, in
 1085 which case BASE performs best.

1086 H.2 Robustness to Mismatched Adversary Assumptions

1087 We evaluate the robustness of our attacks and the baseline MIAs in the scenario where the adversary
 1088 does not have perfect knowledge of the target model architecture and training procedure. In particular,

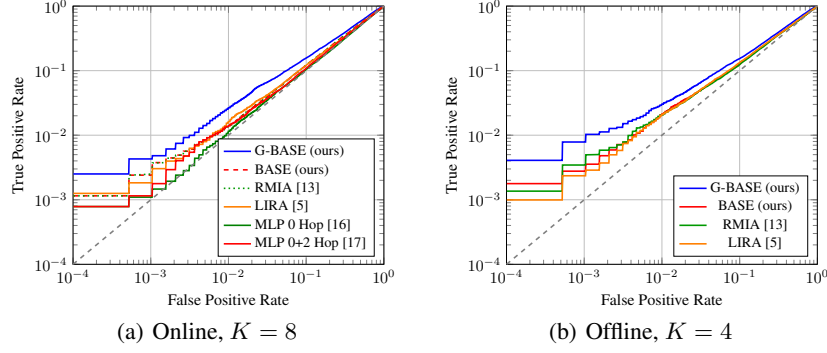


Figure 4: Average ROC curves (10 runs) for the Amazon-Photo dataset with GraphSAGE as target model.

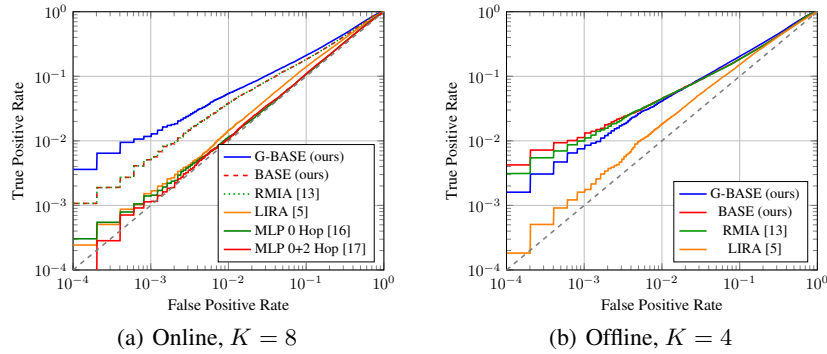


Figure 5: Average ROC curves (10 runs) for the PubMed dataset with GCN as target model.

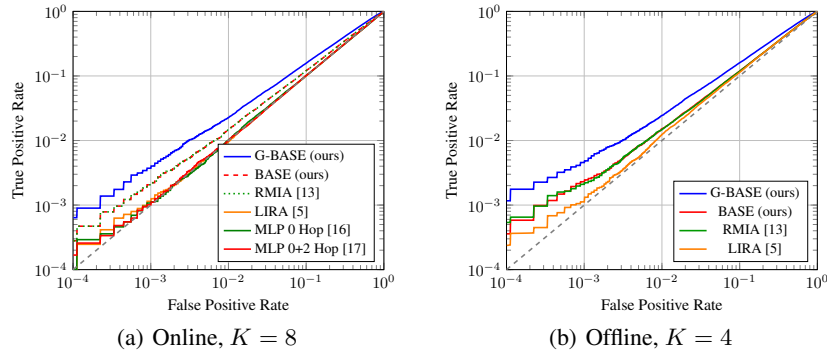


Figure 6: Average ROC curves (10 runs) for the Flickr dataset with GraphSAGE as target model.

we run the attack using shadow models of a different architecture than the target model, trained using a different optimizer (with hyperparameters tuned for the respective model and optimizer). Specifically, the target models are trained using SGD with momentum, whereas the shadow models are trained using Adam. Moreover, the target model is only trained on 35% of the dataset, whereas the adversary trains the shadow models on 50% of the data by means of the usual shadow model training procedure outlined in Algorithm 1. All models are 2-layer GNNs with the final embedding having as many dimensions as there are classes.

Table 8 shows the attack performance over three different datasets. The shadow model architectures are GAT (with 4 and 2 attention heads in the first and second layer, respectively), GraphSAGE with max aggregation, and GCN, on Cora, Citeseer and Pubmed, respectively.

We observe a decline in attack performance across all attacks compared to the ideal setting (see Table 1). LIRA suffers the most and is not competitive with our attacks BASE and G-BASE.

Table 8: Evaluation of attack performance in the case of a distribution shift in the shadow models. Specifically, the adversary uses a model architecture and training procedure that differs from that of the challenger. Performance is measured in terms of AUC and TPR at 1% and 0.1% FPR. The result is reported as the sample mean \pm the standard deviation over 10 random target models and samples of target nodes. The parameter K denotes the number of shadow models. All attacks sees a decline in performance compared to the ideal setting where adversary uses the same model architecture and training procedure as the challenger. Our attacks BASE and G-BASE achieves top performance also in this setting.

| K | ATTACK | CORA (GCN) | | | CITSEER (GAT) | | | PUBMED (GRAPHSAGE) | | |
|-----|---------------------|------------------------------------|-----------------------------------|-----------------------------------|------------------------------------|------------------------------------|-----------------------------------|------------------------------------|-----------------------------------|-----------------------------------|
| | | AUC (%) | TPR@FPR (%) | | AUC (%) | TPR@FPR (%) | | AUC (%) | TPR@FPR (%) | |
| | | | 1% | 0.1% | | 1% | 0.1% | | 1% | 0.1% |
| 8 | MLP (0-HOP) | 57.89 \pm 2.23 | 1.95 \pm 0.86 | 0.40 \pm 0.40 | 63.99 \pm 2.20 | 2.70 \pm 0.94 | 0.20 \pm 0.22 | 50.76 \pm 0.66 | 1.00 \pm 0.22 | 0.11 \pm 0.06 |
| | MLP (0+2-HOP) | 54.50 \pm 1.39 | 1.21 \pm 0.40 | 0.25 \pm 0.21 | 62.48 \pm 1.70 | 2.11 \pm 0.85 | 0.41 \pm 0.50 | 50.74 \pm 0.46 | 1.06 \pm 0.13 | 0.11 \pm 0.07 |
| | LiRA | 62.02 \pm 1.57 | 1.39 \pm 0.68 | 0.21 \pm 0.25 | 64.47 \pm 0.98 | 1.77 \pm 0.67 | 0.18 \pm 0.21 | 52.00 \pm 0.32 | 1.06 \pm 0.12 | 0.11 \pm 0.07 |
| | RMIA | 71.51 \pm 1.13 | 3.81 \pm 0.77 | 0.64 \pm 0.61 | 74.60 \pm 0.88 | 6.63 \pm 1.48 | 0.63 \pm 0.43 | 55.54 \pm 0.56 | 1.70 \pm 0.26 | 0.21 \pm 0.10 |
| | BASE | 71.51 \pm 1.13 | 3.83 \pm 0.79 | 0.64 \pm 0.61 | 74.60 \pm 0.88 | 6.63 \pm 1.48 | 0.65 \pm 0.47 | 55.54 \pm 0.56 | 1.70 \pm 0.26 | 0.21 \pm 0.10 |
| | G-BASE | 67.92 \pm 1.69 | 3.86 \pm 1.90 | 0.55 \pm 0.71 | 69.28 \pm 0.99 | 8.34 \pm 2.10 | 2.17 \pm 2.02 | 57.92 \pm 0.54 | 2.54 \pm 0.26 | 0.39 \pm 0.12 |
| 4 | LiRA (OFF) | 66.65 \pm 0.84 | 1.89 \pm 0.66 | 0.25 \pm 0.33 | 72.86 \pm 1.49 | 2.11 \pm 0.68 | 0.11 \pm 0.22 | 52.73 \pm 0.62 | 1.52 \pm 0.38 | 0.12 \pm 0.05 |
| | RMIA (OFF) | 69.16 \pm 1.08 | 4.62 \pm 1.49 | 0.49 \pm 0.34 | 74.84 \pm 0.93 | 13.50 \pm 1.53 | 4.20 \pm 1.81 | 54.96 \pm 0.65 | 1.90 \pm 0.26 | 0.22 \pm 0.08 |
| | BASE (OFF) | 70.84 \pm 1.07 | 6.34 \pm 1.81 | 1.11 \pm 0.80 | 75.73 \pm 0.92 | 13.06 \pm 1.81 | 4.36 \pm 1.36 | 55.65 \pm 0.61 | 2.37 \pm 0.39 | 0.44 \pm 0.22 |
| | G-BASE (OFF) | 68.47 \pm 0.71 | 3.91 \pm 1.54 | 0.50 \pm 0.30 | 69.93 \pm 1.05 | 7.88 \pm 1.74 | 1.29 \pm 1.64 | 58.26 \pm 0.39 | 2.94 \pm 0.39 | 0.56 \pm 0.21 |

Table 9: Comparison of different sampling strategies for G-BASE. Results are reported as mean \pm standard deviation over 10 different target models and sets of target nodes. The model-independent sampling (MI) and 0-hop MIA sampling (MIA) is performing slightly better than the Metropolis-Hastings (MCMC) sampling strategy. The attack performance is typically not very sensitive to the particular sampling strategy, suggesting that...

| ATTACK | PUBMED (GCN) | | | PUBMED (GAT) | | | PUBMED (GRAPHSAGE) | | |
|---------------|------------------------------------|-----------------------------------|-----------------------------------|------------------------------------|-----------------------------------|-----------------------------------|------------------------------------|-----------------------------------|-----------------------------------|
| | AUC (%) | TPR@FPR (%) | | AUC (%) | TPR@FPR (%) | | AUC (%) | TPR@FPR (%) | |
| | | 1% | 0.1% | | 1% | 0.1% | | 1% | 0.1% |
| G-BASE (MI) | 61.90 \pm 0.43 | 4.56 \pm 0.53 | 0.38 \pm 0.22 | 60.52 \pm 0.60 | 4.31 \pm 0.31 | 0.85 \pm 0.30 | 62.96 \pm 0.42 | 5.22 \pm 0.50 | 1.07 \pm 0.51 |
| G-BASE (MIA) | 61.78 \pm 0.47 | 5.42 \pm 0.42 | 1.17 \pm 0.37 | 60.52 \pm 0.68 | 4.55 \pm 0.41 | 0.98 \pm 0.29 | 62.95 \pm 0.52 | 5.24 \pm 0.43 | 1.16 \pm 0.29 |
| G-BASE (MCMC) | 60.51 \pm 0.32 | 4.00 \pm 0.32 | 0.40 \pm 0.27 | 59.75 \pm 0.70 | 4.07 \pm 0.47 | 0.71 \pm 0.29 | 62.01 \pm 0.51 | 4.83 \pm 0.31 | 1.02 \pm 0.27 |
| ATTACK | FLICKR (GCN) | | | FLICKR (GAT) | | | FLICKR (GRAPHSAGE) | | |
| | AUC (%) | TPR@FPR (%) | | AUC (%) | TPR@FPR (%) | | AUC (%) | TPR@FPR (%) | |
| | | 1% | 0.1% | | 1% | 0.1% | | 1% | 0.1% |
| G-BASE (MI) | 60.46 \pm 0.94 | 2.97 \pm 0.33 | 0.46 \pm 0.18 | 59.58 \pm 0.84 | 3.24 \pm 0.34 | 0.65 \pm 0.21 | 57.24 \pm 1.16 | 2.27 \pm 0.41 | 0.37 \pm 0.11 |
| G-BASE (MIA) | 57.67 \pm 0.62 | 2.52 \pm 0.23 | 0.41 \pm 0.10 | 58.74 \pm 0.85 | 2.96 \pm 0.27 | 0.54 \pm 0.20 | 56.88 \pm 1.11 | 2.20 \pm 0.36 | 0.28 \pm 0.12 |
| G-BASE (MCMC) | 57.47 \pm 0.85 | 2.08 \pm 0.17 | 0.21 \pm 0.07 | 57.19 \pm 0.86 | 2.49 \pm 0.29 | 0.46 \pm 0.17 | 55.26 \pm 0.92 | 1.71 \pm 0.27 | 0.27 \pm 0.10 |
| ATTACK | AMAZON-PHOTO (GCN) | | | AMAZON-PHOTO (GAT) | | | AMAZON-PHOTO (GRAPHSAGE) | | |
| | AUC (%) | TPR@FPR (%) | | AUC (%) | TPR@FPR (%) | | AUC (%) | TPR@FPR (%) | |
| | | 1% | 0.1% | | 1% | 0.1% | | 1% | 0.1% |
| G-BASE (MI) | 57.40 \pm 0.64 | 2.39 \pm 0.72 | 0.24 \pm 0.12 | 56.52 \pm 0.61 | 3.94 \pm 0.73 | 0.61 \pm 0.30 | 56.40 \pm 1.89 | 2.66 \pm 0.73 | 0.43 \pm 0.19 |
| G-BASE (MIA) | 56.05 \pm 1.05 | 2.09 \pm 0.52 | 0.22 \pm 0.12 | 56.74 \pm 0.47 | 3.93 \pm 0.99 | 0.89 \pm 0.54 | 55.22 \pm 1.86 | 1.92 \pm 0.46 | 0.19 \pm 0.13 |
| G-BASE (MCMC) | 55.95 \pm 0.78 | 2.11 \pm 0.40 | 0.28 \pm 0.19 | 55.48 \pm 0.86 | 3.06 \pm 0.77 | 0.72 \pm 0.30 | 55.52 \pm 1.35 | 2.25 \pm 0.46 | 0.29 \pm 0.19 |

H.3 Comparison of Sampling Strategies for G-BASE

In Section 4.3 and Appendix B.2, we proposed three different sampling strategies to sample from $P(\tilde{\mathcal{M}}|\theta, \mathcal{G})$, for the purpose of evaluating the expected value in the Bayes-optimal membership inference rule. Here, we evaluate and compare the impact of the sampling strategy on the attack performance of G-BASE. Table 9 shows the attack performance of G-BASE over different datasets and model architectures, using each of our three sampling strategies; model-independent sampling (MI), 0-hop MIA sampling (MIA), and the Metropolis-Hastings sampling (MCMC). We use BASE to obtain membership probabilities for the 0-hop MIA sampling. The Metropolis-Hastings sampling performs slightly worse than the model-independent sampling or MIA sampling in essentially all cases. Compared to the other sampling strategies, it is more computationally demanding, and requires choosing how many iterations to run between each sample and the length of the burn-in period. We choose a burn-in period of 1000 iterations and 500 iterations between each sample, parameters which could potentially be tuned for each setting to yield better performance. However, since the model-independent sampling performs comparatively well despite disregarding the target model, the loss based signal is likely not as sensitive to the distribution of $\tilde{\mathcal{M}}$. Recall that we have approximated the model distribution $p(\phi|\tilde{\mathcal{M}}, \mathcal{G})$ by a prior distribution $p(\phi|\mathcal{G})$. If $p(\phi|\tilde{\mathcal{M}}, \mathcal{G})$ were estimated more accurately, e.g., by training shadow models on the nodes indicated by $\tilde{\mathcal{M}}$, then the fidelity of the sampling strategy may have a larger impact on the attack performance. In the case of G-BASE, we recommend using either model-independent sampling or 0-hop MIA sampling, since they yield comparative attack performance and are both efficient. Note that for the 0-hop MIA sampling strategy, the same shadow models used for BASE (to obtain membership probabilities for this sampling strategy), can be used for G-BASE, and the membership probabilities can be computed once before running the attack.

Table 10: Comparison of different attacks on Wide ResNet-28-2 trained on CIFAR-10 and CIFAR-100. Performance is measured in terms of AUC and TPR at 1% and 0.1% FPR.

| K | ATTACK | CIFAR-10 | | | CIFAR-100 | | |
|----|-------------|---------------------|--------------------|--------------------|---------------------|---------------------|--------------------|
| | | AUC | TPR@FPR (%) | | AUC | TPR@FPR (%) | |
| | | | 1% | 0.1% | | 1% | 0.1% |
| 32 | BASE | 62.94 ± 2.06 | 5.92 ± 1.27 | 1.66 ± 0.43 | 74.80 ± 3.22 | 12.08 ± 3.35 | 4.11 ± 1.69 |
| | RMIA | 62.94 ± 2.06 | 5.92 ± 1.27 | 1.66 ± 0.43 | 74.80 ± 3.22 | 12.08 ± 3.35 | 4.11 ± 1.69 |
| | LIRA | 61.00 ± 1.45 | 5.21 ± 0.91 | 1.32 ± 0.33 | 72.65 ± 1.80 | 9.99 ± 1.86 | 2.47 ± 0.87 |
| 16 | BASE (OFF) | 61.67 ± 2.10 | 5.67 ± 1.29 | 1.64 ± 0.39 | 71.35 ± 3.36 | 8.89 ± 2.97 | 3.07 ± 1.42 |
| | RMIA (OFF) | 62.45 ± 2.10 | 5.61 ± 1.23 | 1.57 ± 0.40 | 70.65 ± 3.30 | 7.35 ± 2.21 | 2.07 ± 0.94 |
| | LIRA (OFF) | 60.51 ± 1.75 | 4.10 ± 0.61 | 0.94 ± 0.25 | 72.33 ± 2.66 | 8.36 ± 1.65 | 1.60 ± 0.69 |
| 8 | BASE | 62.64 ± 1.96 | 5.21 ± 0.99 | 1.19 ± 0.38 | 73.98 ± 3.18 | 10.34 ± 2.91 | 2.81 ± 0.95 |
| | RMIA | 62.64 ± 1.96 | 5.21 ± 0.99 | 1.19 ± 0.38 | 73.98 ± 3.18 | 10.34 ± 2.91 | 2.81 ± 0.95 |
| | LIRA | 58.79 ± 0.99 | 3.64 ± 0.67 | 0.85 ± 0.25 | 69.22 ± 1.60 | 7.29 ± 1.37 | 1.60 ± 0.58 |
| 4 | BASE (OFF) | 61.54 ± 1.98 | 5.05 ± 1.06 | 1.23 ± 0.39 | 71.26 ± 3.28 | 8.23 ± 2.62 | 2.45 ± 1.20 |
| | RMIA (OFF) | 62.17 ± 1.96 | 4.94 ± 1.07 | 1.19 ± 0.39 | 70.48 ± 3.25 | 6.45 ± 1.88 | 1.45 ± 0.67 |
| | LIRA (OFF) | 59.80 ± 1.67 | 3.77 ± 0.76 | 0.78 ± 0.24 | 71.00 ± 2.68 | 7.27 ± 1.45 | 1.26 ± 0.57 |

H.4 i.i.d. Data

To demonstrate BASE on i.i.d. data, we train a Wide ResNet [37] with depth 28 and width 2 on both CIFAR-10 and CIFAR-100 for 100 epochs using standard data augmentations and early stopping. Each experiment is averaged across ten target models. For CIFAR-10, the target model achieves a mean training accuracy of $93.73\% \pm 1.15\%$ and a test accuracy of $80.45\% \pm 1.53\%$. For CIFAR-100, the model reaches a training accuracy of $75.97\% \pm 6.68\%$ and a test accuracy of $49.89\% \pm 1.88\%$.

Table 10 presents the attack performance of BASE, RMIA, and LIRA. In the online setting, BASE and RMIA exhibit identical performance, in line with Theorem 2. Both methods consistently outperform LIRA across the evaluated configurations. In the offline setting, while the differences are less pronounced, BASE achieves superior performance over both RMIA and LIRA at low false positive rates.