BAYESIAN MEMBERSHIP PRIVACY FOR GRAPH NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing privacy analyses for Graph Neural Networks (GNNs) largely inherit assumptions from non-graph settings, overlooking structural correlations and stochastic training-graph sampling. In particular, node-dependent priors make type-I and type-II errors alone insufficient to characterize the best membership inference test. To address this, we introduce Bayesian Membership Privacy (BMP), a sampling-aware formulation of node-level membership privacy that incorporates node-dependent priors and treats graph sampling probabilities as part of the adversary's knowledge. BMP casts membership inference as a Bayesian hypothesis test and accordingly quantifies membership privacy in terms of posterior membership probability. We explore theoretical properties of BMP in relation to the existing definitions in the literature. We further propose a practical, sampling-aware auditing mechanism to estimate the parameters of BMP as a measure of node-level privacy leakage in GNNs. We conduct experiments on benchmark graph datasets and show that BMP yields fine-grained privacy insights that are not visible through global attack accuracy alone.

1 INTRODUCTION

Graph Neural Networks (GNNs) are now a standard modeling paradigm for learning from structured and relational data, with successful applications in various domains. Alongside their increasing adoption, recent studies have shown that trained GNNs can inadvertently expose information about the data used during training, raising concerns about privacy in graph-based learning (Wang & Wang, 2024; Wu et al., 2021; Dai et al., 2022). Existing membership-inference analyses, however, largely inherit a view developed for independent records: membership is treated as the same binary inference problem for every example.

We study node-level membership privacy in GNNs trained for node classification, where the adversary seeks to determine whether a specific node was included in the training set. Unlike i.i.d. settings, nodes in a graph are interdependent through edges and overlapping neighborhoods. As a result, the inclusion of a single node can affect the learned representations and predictions of many others. Empirical implications of this phenomenon on MI were first observed in (Olatunji et al., 2021), where membership inference risk did not consistently correlate with the generalization gap.

Structural dependence also limits what conventional, uniform analyses reveal about node-level privacy. In particular, node-level DP (Sajadmanesh et al., 2023; Olatunji et al., 2023; He et al., 2025b) typically assumes a worst-case adversary who knows the entire dataset except for one individual. In graph settings, however, nodes are coupled through edges and shared neighborhoods, and training nodes are selected via a sampling process that interacts with this structure. As a result, some nodes are inherently more likely to be included than others, and knowledge of the remaining graph can already reveal substantial information about a node's membership. Moreover, in i.i.d. settings, membership-inference performance can be formally related to DP guarantees under exchangeability assumptions. No analogous relationship is known for GNNs because structural dependencies break exchangeability between training and non-training nodes (Khosla, 2026).

We address this gap by a new privacy notion called *Bayesian Membership Privacy* (BMP), which models node-level membership inference as Bayesian hypothesis testing with node-dependent priors. BMP accounts for heterogeneity introduced by the training-set sampling process and bounds the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

adversary's posterior membership belief after observing the trained model. This shifts privacy analysis from a single population-level estimate to node-specific guarantees, revealing risks that global attack metrics can conceal. Our key contributions are as follows.

- **Sampling-aware membership privacy.** BMP is motivated by non-uniform node sampling in graphs but is *applicable more broadly whenever records have non-uniform inclusion probabilities*, for example, when rare classes are oversampled, users are selected through geographic clusters, or high-importance records are preferentially sampled. BMP incorporates node-specific prior inclusion probabilities and allows asymmetric protection of membership and non-membership, thereby capturing privacy heterogeneity induced by sampling (Section 3).
- **Theoretical characterization.** We establish prior-induced lower bounds, post-processing and composition properties, and connections to likelihood-ratio membership privacy and Pufferfish-style notions. We then characterize BMP through lower bounds on the type-I and type-II errors and the expected decision cost of arbitrary membership-inference attacks. (Section 3.1) Differently from its DP counterparts, the feasible error region of type-I and type-II errors is asymmetric and depends on node's prior inclusion probability. (See Figure 1.)
- **Auditing of BMP.** We make BMP operational through an end-to-end Bayesian auditing framework that translates membership-inference observations from a finite number of reusable shadow models into posterior distributions over BMP parameters. The framework incorporates heterogeneous membership priors and quantifies uncertainty from limited attack evidence, yielding uncertainty-aware estimates of aggregate BMP guarantees (see Section 4).
- **Empirical evidence of hidden privacy heterogeneity.** We instantiate BMP auditing on benchmark graph datasets, varying the sampling strategy, training-set size, edge structure, GNN architecture (including differential private GNN), and noise level. BMP auditing provides fine-grained, uncertainty-aware privacy estimates and reveals sampling-dependent variation that global membership-privacy measures fail to capture, including for differentially private GNNs.

2 Background and Related Work

Graph Neural Networks and Training-Graph Membership. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote a graph with $N = |\mathcal{V}|$ nodes, where $\mathcal{V} = \{1, \dots, N\}$ and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the edge set. Each node $i \in \mathcal{V}$ has a feature vector $\mathbf{x}_i \in \mathbb{R}^d$, collected in the feature matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$. For a c-class node-classification task, let $\mathbf{y}_i \in \{0,1\}^c$ denote the one-hot label of node i, and let $\mathbf{Y}_{\text{train}}$ denote the labels available for the training nodes. A GNN with L message-passing layers computes node representations by iteratively aggregating information from local neighborhoods. Let $\mathcal{N}(i) = \{j \in \mathcal{V} : (i,j) \in \mathcal{E}\}$ denote the one-hop neighborhood of node i, and let $\mathbf{x}_i^{(0)} = \mathbf{x}_i$. At layer ℓ,

$$\mathbf{z}_i^{(\ell)} = \mathrm{AGG}^{(\ell)}(\mathbf{x}_i^{(\ell-1)}, \{\mathbf{x}_j^{(\ell-1)} : j \in \mathcal{N}(i)\}), \qquad \mathbf{x}_i^{(\ell)} = \mathrm{TRANS}^{(\ell)}(\mathbf{z}_i^{(\ell)}),$$

where $\mathrm{AGG}^{(\ell)}$ and $\mathrm{TRANS}^{(\ell)}$ depend on the chosen GNN architecture. The final representation is mapped to class probabilities, $\hat{\mathbf{y}}_i = \mathrm{softmax}(\phi_W(\mathcal{G}, \mathbf{X})_i) \in \Delta^{c-1}$, where ϕ_W denotes the trained GNN with parameters $W \in \mathcal{W}$ and Δ^{c-1} is the probability simplex. Because message passing propagates information across edges, $\hat{\mathbf{y}}_i$ generally depends on the features of nodes in the L-hop neighborhood of i. Details of the GNN architectures used in our experiments are provided in Appendix A.9.

This neighborhood dependence is important for defining membership. In the *transductive* setting, the full graph $\mathcal{G}$ and feature matrix $\mathbf{X}$ are available during training, while only the supervision set varies. Hence all nodes already participate in the model input, making node-level membership ambiguous for the privacy notion studied here. We therefore, consistent with prior works, focus on the *inductive* setting, where training is performed on a subgraph, $\mathcal{G}_{\text{train}} = (\mathcal{V}_{\text{train}}, \mathcal{E}_{\text{train}}) \subseteq \mathcal{G}$, sampled from an underlying graph, and the learned model is subsequently applied to unseen nodes or graphs. Specifically, the sampling process may assign different prior probabilities of training-set inclusion to different nodes, depending on graph structure and the construction of the training graph. BMP explicitly accounts for this sampling-induced heterogeneity when defining node-level membership privacy, rather than assuming a common prior for all nodes.

Membership Privacy and Inference Attacks. Membership privacy concerns whether an adversary can infer if a particular record was used for training (Li et al., 2013). Membership Inference At-

tacks (MIAs) operationalize this risk by testing whether model outputs distinguish members from non-members (Shokri et al., 2017). Prior work has linked such leakage to overfitting and model confidence and developed increasingly strong black-box attacks (Yeom et al., 2018; Salem et al., 2018; Carlini et al., 2022). For GNNs, MIAs have likewise demonstrated leakage about nodes, edges, and subgraphs (Olatunji et al., 2021; He et al., 2021; Wang et al., 2021). For independently sampled data, MIA performance can lower-bound the privacy budget of DP mechanisms because exchangeability ensures that membership distinguishability arises from the learning algorithm (Jayaraman & Evans, 2019; Humphries et al., 2023). This symmetry fails for graphs, motivating a new privacy notion.

Privacy Preserving GNNs. Existing approaches to privacy-preserving GNN training broadly rely on knowledge distillation (Olatunji et al., 2023), DP-SGD-style gradient perturbation (Daigavane et al., 2022; Xiang et al., 2024), or noise added during neighborhood aggregation (Sajadmanesh et al., 2023). Recent methods further improve the privacy–utility trade-off through decoupled propagation, progressive perturbation, and personalized or local privacy mechanisms (Zhang et al., 2024; Sajadmanesh & Gatica-Perez, 2024; He et al., 2025a; Lei et al., 2025). Our work is complementary: rather than designing a private GNN, we audit a given mechanism under BMP. BMP provides a valid formal linkage between MIA error rates and privacy guarantees in graph settings.

3 Bayesian Membership Privacy

Setting. Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with node features $\mathbf{X}$ and labels $\mathbf{Y}$, we consider a randomized learning algorithm operating in two stages. First, a subset $S \subseteq \mathcal{V}$ is sampled using a stochastic subset selection mechanism, inducing a training subgraph $\mathcal{G}_S$ and $\mathbf{X}_S, \mathbf{Y}_S$ accordingly. Second, a (randomized) training procedure Φ is applied to $(\mathcal{G}_S, \mathbf{X}_S, \mathbf{Y}_S)$ to produce an output $W \in \mathcal{W}$.

Let $\mathcal{P}$ represent *the adversary's prior distribution* over the sampling mechanism, conditional on their pre-release knowledge. Crucially, $\mathcal{P}$ need not coincide with the actual selection mechanism, say $\mathcal{P}_*$, which may depend on some sensitive information $\mathbf{Z}$ available to the data holder but not to the adversary. In that case, $\mathcal{P}$ is obtained by marginalizing out $\mathcal{P}_*$ over the uncertainty about $\mathbf{Z}$,

$$\mathcal{P}(S = s|\mathcal{G}, \mathbf{X}, \mathbf{Y}) = \sum_{\mathbf{z}} \mathcal{P}_*(S = s \mid \mathcal{G}, \mathbf{X}, \mathbf{Y}, \mathbf{Z} = \mathbf{z}) \Pr(\mathbf{Z} = \mathbf{z}|\mathcal{G}, \mathbf{X}, \mathbf{Y}).$$

We define the overall learning algorithm $\mathcal{A} = (\mathcal{P}, \Phi)$ from the adversary's point of view. The adversary knows the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the node features and labels $\mathbf{X}$, $\mathbf{Y}$, and the algorithm $\mathcal{A}$, and the algorithm's output W. However, the adversary does *not* know the sampled subset S. Membership status can itself reveal sensitive information, even when the graph and node attributes are known. Specifically, when the adversary has incomplete knowledge of the sampling mechanism ($\mathcal{P} \neq \mathcal{P}_*$), membership or non-membership may reveal hidden selection criteria, such as vulnerability or promotion potential. Even when the sampling mechanism is fully known ($\mathcal{P} = \mathcal{P}_*$), membership may reveal sensitive treatment received *after* selection, such as genome sequencing. Appendix A.2 provides three concrete examples of these cases.

Probabilistic Model: Membership prior, likelihood, and posterior. We construct the probabilistic model corresponding to the setup above *with respect to the adversary's knowledge*. For simplicity, we will use $\Pr(\cdot)$ without explicitly showing conditioning on $\mathcal{G}, \mathbf{X}, \mathbf{Y}$.

We begin with *prior membership probabilities*. For each node $i \in \{1, \ldots, N\}$, we define the membership variable $M_i := \mathbb{I}(i \in S) \in \{0, 1\}$ and let γ_i denote the marginal distribution of M_i,

$$\gamma_i(m) := \Pr(M_i = m), \quad m \in \{0, 1\}.$$

We assume $0 < \gamma_i(1) < 1$ for every node considered. Next, we define *membership likelihood* given the output. We assume that, the conditional distribution of W given $M_{1:N} = m_{1:N} \in \{0, 1\}^N$ has a probability density $f(\cdot|m_{1:N})$. This allows us to define the node-level membership likelihood

$$f_i(w|m_i) := \sum_{m_{-i}} f(w|m_{1:N}) \Pr(M_{-i} = m_{-i}|M_i = m_i).$$

Finally, we define the *posterior membership probabilities*. For any $i \in \{1, \ldots, N\}$ and $m_i \in \{0, 1\}$, and $w \in \mathcal{W}$ such that $f_0(w) := f_i(w|1) \Pr(M_i = 1) + f_i(w|0) \Pr(M_i = 0) > 0$, define marginal posteriors conditional on $\mathcal{A}$'s output

$$\pi_i(m_i|w) = \Pr(M_i = m_i|W = w) = \gamma_i(m_i) f_i(w|m_i)/f_0(w).$$

Definitions of Membership Privacy. We begin with a non-Bayesian definition of membership privacy that focuses on the likelihood ratio given the membership information of a single node, and it can be viewed as a pufferfish-type privacy (Kifer & Machanavajjhala, 2014) specific to membership.

Definition 1 (Membership privacy (MP)). *We say $\mathcal{A}$ is ε-MP if for any $w \in \mathcal{W}$ such that $f_0(w) > 0$, and for any $i \in \{1, \ldots, N\}$, we have $e^{-\varepsilon} \leq f_i(w|1)/f_i(w|0) \leq e^{\varepsilon}$.*

MP is likelihood-based and abstracts away prior membership probabilities. In relational data, however, prior inclusion probabilities are shaped by the training-graph sampling process and may themselves encode information relevant to privacy. Thus, leakage induced by structural dependencies is not fully captured by MP. We, therefore, adopt a Bayesian perspective, where privacy is defined relative to explicit inclusion priors that serve as a proxy for the data-generating process.

Definition 2 (Bayesian Membership Privacy (BMP)). *We say $\mathcal{A}$ is $(\varepsilon_L, \varepsilon_R)$-BMP if any $w \in \mathcal{W}$ such that $f_0(w) > 0$, and for any $i \in \{1, \ldots, N\}$, we have*

$$e^{-\varepsilon_L} \leq \pi_i(1|w)/\pi_i(0|w) = [\gamma_i(1) f_i(w|1)]/[\gamma_i(0) f_i(w|0)] \leq e^{\varepsilon_R}. \tag{1}$$

The BMP-R and BMP-L bear similarities with the positive posterior membership privacy of Li et al. (2013), which we discuss in Section A.3. Apart from the inclusion of the prior membership probabilities, another feature of Definition 2 is that the inequality (1) is *asymmetric*. This is deliberately done for flexibility and to be able to address certain issues regarding the meaning of privacy. The most important of those issues is the significance of the knowledge of membership relative to the knowledge of non-membership. In many cases, knowledge of non-membership is not a privacy breach, and one should only focus on the knowledge of membership by the adversary. The following one-sided versions of BMP correspond to the two extreme cases of such an imbalance.

Definition 3 (BMP-R, BMP-L). *$\mathcal{A}$ is called ε-BMP-R if it is (∞, ε)-BMP, Similarly, $\mathcal{A}$ is called ε-BMP-L if it is (ε, ∞)-BMP.*

The membership privacy definitions given above focus on marginal (unconditional) posterior membership probabilities. This is in contrast with classical DP definitions, which limit the attacker's ability to discover the membership of a single node given the membership information of every other node. For reference, we provide the DP versions of MP and BMP definitions in Appendix A.4. While the stronger DP guarantees are attractive in classical i.i.d. settings, they can be overly restrictive in graph-structured domains where memberships are strongly dependent. Enforcing worst-case bounds over arbitrary membership configurations may either severely degrade utility.

3.1 Properties of BMP

We focus on BMP in Definition 2 and study its properties. The proofs of the propositions are in Appendix A.5. First, we show that the BMP parameters $\varepsilon_L, \varepsilon_R$ are lower-bounded by the ratio of priors. Let $\eta_i := \gamma_i(1)/\gamma_i(0)$ and $\eta_{\max} := \max\{\eta_1, \ldots, \eta_N\}$ and $\eta_{\min} = \min\{\eta_1, \ldots, \eta_N\}$.

Proposition 1 (Bounds by membership prior). *If $\mathcal{A}$ is $(\varepsilon_L, \varepsilon_R)$-BMP, $e^{-\varepsilon_L} \leq \eta_i \leq e^{\varepsilon_R}$ for all $i \in \{1, \ldots, N\}$. That is, $\varepsilon_R \geq \log \eta_{\max}$ and $\varepsilon_L \geq -\log \eta_{\min}$.*

This proposition highlights the importance of the sampling mechanism in the BMP definition. Indeed, even a completely random training algorithm, for example, one that discards all input and outputs pure noise, cannot undo leakage that arises solely from the sampling process.

BMP satisfies the post-processing property, making it suitable for complex pipelines where model outputs might be reused, e.g. for generating explanations, or transformed, e.g., for ranking nodes.

Proposition 2 (Post-processing). *Let $\varphi : \mathcal{W} \to \mathcal{Z}$ be a measurable function. If $\mathcal{A} = (\mathcal{P}, \Phi)$ is $(\varepsilon_L, \varepsilon_R)$-BMP, then $\mathcal{B} = (\mathcal{P}, \Phi')$ where $\Phi' = \varphi \circ \Phi$ is $(\varepsilon_L, \varepsilon_R)$-BMP, too.*

A relation between MP and BMP can also be established.

Proposition 3 (Relation between MP and BMP). *If $\mathcal{A}$ is ε-MP, then it is also $(\varepsilon_L, \varepsilon_R)$-BMP with $\varepsilon_L = \varepsilon - \log \eta_{\min}, \quad \varepsilon_R = \varepsilon + \log \eta_{\max}$. Furthermore, if $\mathcal{A}$ is $(\varepsilon_L, \varepsilon_R)$-BMP, it is also ε-MP where $\varepsilon = \max\{\varepsilon_L + \log \eta_{\max}, \varepsilon_R - \log \eta_{\min}\}$.*

A composition of BMP algorithms exists when the sampling mechanisms are independent under the adversary's *a priori* knowledge. See Proposition 6 in Appendix A.5 for a formal statement.

Statistical Guarantees of BMP. We analyze the performance of any MIA that uses the output of $\mathcal{A}$ as its input. First, we formalize an MIA related to $\mathcal{A}$.

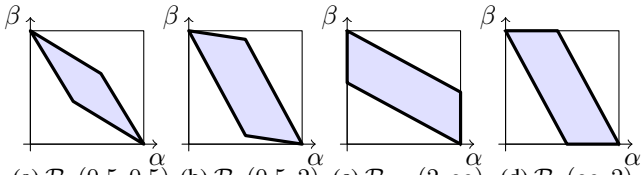
Definition 4 (MIA – Hypothesis Testing Perspective). *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with node feature matrix $\mathbf{X}$ and label matrix $\mathbf{Y}$. Let $\mathcal{A} = (\mathcal{P}, \Phi)$ be a (possibly randomized) algorithm operating on $(\mathcal{G}, \mathbf{X}, \mathbf{Y})$, producing an output $\hat{Y} \in \mathcal{W}$. For $i \in \{1, \dots, N\}$, a MIA, denoted as $\mathcal{M}(i, D, \mathcal{A})$, is a hypothesis test for the hypotheses $H_0 : M_i = 0$ vs $H_1 : M_i = 1$, whose decision rule $D : \mathcal{W} \mapsto \{0, 1\}$ maps the output of $\mathcal{A}$ to a binary decision in $\{0, 1\}$.*

For a MIA $\mathcal{M}(i, D, \mathcal{A})$, define its type-I error probability $\alpha_i = \Pr(D(W) = 1 | M_i = 0)$ and type-II error probability as $\beta_i = \Pr(D(W) = 0 | M_i = 1)$. We have the following crucial result for (α_i, β_i).

Theorem 4 (BMP and error probabilities of MIA)**.** *$\mathcal{A}$ is $(\varepsilon_L, \varepsilon_R)$-BMP if and only if, for any decision rule D and any $i \in \{1, \dots, N\}$, the MIA $\mathcal{M}(i, D, \mathcal{A})$ satisfies $(\alpha_i, \beta_i) \in \mathcal{R}_{\eta_i}(\varepsilon_L, \varepsilon_R)$, where for $\varepsilon_L, \varepsilon_R \in \mathbb{R}$ and $\eta \geq 0$,*

$$\mathcal{R}_\eta(\varepsilon_L, \varepsilon_R) = \left\{(\alpha, \beta) \in [0,1]^2 : \alpha + \eta e^{\varepsilon_L} \beta \in [1, \eta e^{\varepsilon_L}], \quad \alpha + \eta e^{-\varepsilon_R} \beta \in [\eta e^{-\varepsilon_R}, 1]\right\}. \tag{2}$$

Figure 1 illustrates $\mathcal{R}_\eta(\varepsilon_L, \varepsilon_R)$ for different choices of $\varepsilon_L, \varepsilon_R, \eta$ and note the allowed asymmetry in the region. The prior-unaware and symmetric version $\mathcal{R}_1(\varepsilon, \varepsilon)$ is well known (Kairouz et al., 2017, Theorem 2.1) and has been used for empirical privacy estimation (Jagielski et al., 2020; Nasr et al., 2021; Zanella-Beguelin et al., 2023; Yıldırım et al., 2026). The definition $\mathcal{R}_\eta(\varepsilon_L, \varepsilon_R)$ in (2) generalizes $\mathcal{R}_1(\varepsilon, \varepsilon)$ significantly, while it is still related to the Neyman-Pearson lemma. More importantly, the one-to-one relation in Theorem 4 will be essential for the method in Section 4 for empirical estimation of the privacy parameters $\varepsilon_L, \varepsilon_R$ of BMP. Note that, differently from the prior-unaware classical DP-related definitions, the allowed region of type-I and type-II error probabilities *varies* across $i \in \{1, \dots, N\}$ due to η_i.



(a) $\mathcal{R}_1(0.5, 0.5)$ (b) $\mathcal{R}_4(0.5, 2)$ (c) $\mathcal{R}_{1/4}(2, \infty)$ (d) $\mathcal{R}_4(\infty, 2)$

Figure 1: Region $\mathcal{R}_\eta(\varepsilon_L, \varepsilon_R)$ for different choices. Panel (a) corresponds to ε-MP, (c) & (d) correspond to ε-BMP-L and ε-BMP-R, respectively

In the Bayesian framework, type I and type II errors are not sufficient alone to characterize the capability of the adversary; one can gain additional insight into the adversary's capability by studying the expected cost of a wrong decision and the probability of a wrong decision as a special case. For $j = 0, 1$, let c_j be the cost of an error under $H_j : M_i = j$. Then, the overall cost can be defined as

$$\text{Cost} = c_0(1 - M_i)D(W) + c_1 M_i(1 - D(W)).$$

The next theorem provides a lower bound on the expected cost of error and, as a special case, the probability of a wrong decision, $D(W) \neq M_i$.

Theorem 5 (BMP and Expected Cost of Error of MIA)**.** *If $\mathcal{A}$ is $(\varepsilon_L, \varepsilon_R)$-BMP, then for any decision rule D and any $i \in \{1, \dots, N\}$, expected cost of the MIA $\mathcal{M}(i, D, \mathcal{A})$ is lower-bounded by*

$$\mathbb{E}(\text{Cost}) \geq \max\left\{ \min\left(\frac{c_0}{1 + e^{\varepsilon_R}}, \frac{c_1}{1 + e^{\varepsilon_L}}\right), \gamma_i(0) \min\left(\frac{c_1}{e^{\varepsilon_L}}, c_0\right), \gamma_i(1) \min\left(\frac{c_0}{e^{\varepsilon_R}}, c_1\right) \right\}. \tag{3}$$

In particular, $\Pr(M_i \neq D) \geq \max\left\{1/(1 + e^{\max\{\varepsilon_L, \varepsilon_R\}}), \gamma_i(0)/\min\{1, e^{\varepsilon_L}\}, \gamma_i(1)/\min\{1, e^{\varepsilon_R}\}\right\}$.

4 Auditing Bayesian Membership Privacy

We audit $\mathcal{A}$ under BMP by inferring $(\varepsilon_L, \varepsilon_R)$ from MIA-based evidence. The audit has three steps: a prior-aware attack that uses shadow models (Section 4.1); a leave-one-out protocol that reuses shadow models to measure its type-I and type-II errors; and a probabilistic model that converts the resulting error counts into posterior distributions over BMP parameters (Section 4.2). The final step adapts Bayesian privacy estimation (Zanella-Beguelin et al., 2023; Yıldırım et al., 2026) to BMP.

4.1 MIA: Design and Measurement Protocol

For a target node ν, let $\mathcal{I}_\nu$ be its graph query input and y_ν its true label. Given a trained GNN ϕ_w with model parameters $w \in \mathcal{W}$, we use the correct-class logit margin $z_\nu(\phi_w) = [\phi_w(\mathcal{I}_\nu)]_{y_\nu} -$

$\log \sum_{y \neq y_\nu} e^{[\phi_w(\mathcal{I}_\nu)]_y}$ as the scalar attack statistic. This statistic is a graph-node analogue of the confidence or loss statistic used by likelihood-ratio membership attacks, including LiRA (Carlini et al., 2022). For each target node, shadow models are partitioned according to whether ν was included in their sampled training graph. We fit distributions for the score under the two hypotheses, $H_0 : M_\nu = 0$ and $H_1 : M_\nu = 1$, using normal approximations $\mathcal{N}(\mu_0, \sigma_0^2)$ and $\mathcal{N}(\mu_1, \sigma_1^2)$. The resulting attack is a prior-aware likelihood-ratio test with threshold parameter $\tau > 0$:

$$\hat{m}_\tau(z_\nu(\phi_w)) = \mathbb{I}\left[\mathcal{N}(z_\nu(\phi_w); \mu_1, \sigma_1^2)/\mathcal{N}(z_\nu(\phi_w); \mu_0, \sigma_0^2) > \tau/\eta_\nu\right]. \tag{4}$$

where $\eta_\nu = \gamma_\nu(1)/\gamma_\nu(0)$. Equation (4) is a LiRA-style likelihood-ratio test with node-specific threshold τ/η_ν. Sweeping τ yields an operating curve for the attack, while the $\tau = 1$ rule is the Bayes-optimal decision rule under equal 0–1 misclassification costs.

Measurement protocol. Algorithm 1 (provided in Appendix A.1) measures the performance of the attack in (4) economically. First, we sample J training graphs $S_{1:J}$ under $\mathcal{P}$ and train GNNs on each graph, obtaining J outputs $W_{1:J}$ and shadow models $\phi_j = \phi_{W_j}$ for $j = 1, \ldots, J$. Then, for each node $\nu_k \in \mathcal{T}$ and model ϕ_j, we form a *challenge* (ν_k, ϕ_j). For this challenge, all the shadow models *except* ϕ_j are used to learn the absence (H_0) and presence (H_1) score distributions of ν_k. A decision is made for the challenge (ν_k, ϕ_j) using the learned hypotheses. In this way, for each node $\nu_k \in \mathcal{T}$ the MIA is tested for $n_{0,k} := \sum_{j=1}^{J} \mathbb{I}(\nu_k \notin S_j)$ times under H_0 and $n_{1,k} = J - n_{0,k}$ times under H_1, and the false-positive and false-negative counts are collected as A_k, B_k. This leave-one-out construction avoids fitting and testing on the same model while reusing the same pool of J trained GNNs. Thus, it requires only J trained shadow models and $\mathcal{O}(KJ)$ attack decisions, rather than training fresh in/out shadow ensembles for every challenge.

Availability of prior membership probabilities. Under certain sampling mechanisms, such as random subsampling, $\gamma_\nu(1)$, hence η_ν, is exactly known. Otherwise, an estimate of $\gamma_\nu(1)$ is needed. The prior membership probabilities can be estimated for all nodes by sampling $J_\gamma \gg 1$ subgraphs independently and calculating the frequency of inclusion in a subgraph. The subgraphs sampled for shadow models can already be used for this purpose, in which case $J_\gamma = J$ and $\eta_\nu(1) \approx n_{1,\nu}/J$. In practice, however, prior estimation can be performed separately and with much larger J_γ, since sampling a subgraph is typically much less demanding than obtaining a shadow model.

4.2 Joint Probabilistic Model and MCMC for estimating BMP parameters

Modeling error probabilities: The key observation underlying our estimator is that Theorem 4 establishes a direct relationship between the privacy parameters $(\varepsilon_L, \varepsilon_R)$ and the admissible type-I/type-II error pairs of any MIA. Consequently, we first model the latent error probabilities (α_k, β_k) associated with attacks targeting individual nodes, and use them in a joint probabilistic model as an intermediate layer connecting privacy parameters to the observed attack statistics.

Suppose we have MIA results for a set of nodes $\mathcal{T} = \{\nu_1, \ldots, \nu_K\} \subseteq \{1, \ldots, N\}$ where $K = |\mathcal{T}|$. For each k, let (α_k, β_k) be the type-I and type-II errors of the designed MIA, $\mathcal{M}(\nu_k, D, \mathcal{A})$. By Theorem 4, if $\mathcal{A}$ is $(\varepsilon_L, \varepsilon_R)$-BMP, then $(\alpha_k, \beta_k) \in \mathcal{R}_{\eta_{\nu_k}}(\varepsilon_L, \varepsilon_R)$ for every MIA and $k = 1, \ldots, K$. Motivated by this support constraint, we adopt the modeling assumption

$$(\alpha_k, \beta_k)|\varepsilon_L, \varepsilon_R \overset{\text{indep}}{\sim} \text{Unif}(\mathcal{R}_{\eta_{\nu_k}}(\varepsilon_L, \varepsilon_R)). \tag{5}$$

Uniformity over the admissible region is a modeling assumption; we discuss an alternative modeling choice in Section A.8. Acknowledging multiple possible choices, we will generically denote the conditional distribution of α_k, β_k given $\varepsilon_L, \varepsilon_R$ by $h_{\eta_k}(\alpha_k, \beta_k|\varepsilon_L, \varepsilon_R)$.

Likelihood model for error counts: The error probabilities (α_k, β_k) are not directly observable; instead, they yield the observed false-positive and false-negative counts obtained from repeated MIA evaluations. For each node k, assume that MIA is applied $n_{0,k}$ times under $H_0 : M_{\nu_k} = 0$ and $n_{1,k}$ times under $H_1 : M_{\nu_k} = 1$, and we obtain the respective counts A_k, B_k of type I and type II errors. The conditional distribution of A_k, B_k can generically be shown to have the probability distribution

$$g_\theta(a_k, b_k|\alpha_k, \beta_k) = \Pr(A_k = a_k, B_k = b_k|\alpha_k, \beta_k), \tag{6}$$

where θ contains parameters (if any) used to construct g_θ, which may be assumed unknown and estimated together with $(\varepsilon_L, \varepsilon_R)$, and we have suppressed $n_{0,k}, n_{1,k}$ for notational simplicity. A simple choice is to model A_k, B_k as independent Binomial random variables, which is the choice

in our experiments in Section 5. This is appropriate when the MIA instances are conditionally approximately independent given (α_k, β_k). We also discuss in Section A.8 an alternative modeling approach that accounts for dependence among the test decisions.

The overall model and sampling from the posterior distribution: Combining the distributions for the error rates and observed error counts, and assuming an uninformative prior $p(\varepsilon_L, \varepsilon_R)$ on $(\varepsilon_L, \varepsilon_R)$, and $p(\theta)$ on θ, we have the joint probability distribution

$$p(\varepsilon_L, \varepsilon_R, \theta, (\alpha, \beta)_{1:K} | (a, b)_{1:K}) \propto p(\varepsilon_L, \varepsilon_R) p(\theta) \cdot \prod_{k=1}^{K} h_{\eta_k}(\alpha_k, \beta_k | \varepsilon_L, \varepsilon_R) g_\theta(a_k, b_k | \alpha_k, \beta_k). \quad (7)$$

The posterior distribution in (7) can be used to generate samples via MCMC. To that end, we provide Algorithm 2 in Appendix A.6, which, being a variant of MHAAR (Andrieu et al., 2020), is a suitable MCMC algorithm for latent variable models such as the one in (7).

Remark 1. The methodology can be modified for the other privacy definitions. For example, for BMP-L (resp. BMP-R), it suffices to fix $\varepsilon_R = \infty$ (resp. $\varepsilon_L = \infty$). For MP, one can enforce $\varepsilon_L = \varepsilon_R$ in the joint probability model and run Algorithm 2 with flat priors, i.e., $\gamma_k(1) = 0.5$ for all k.

Extension of the methodology. There are two main potential limitations to the estimation method: (i) The independence (across nodes) and uniformity of error rates can be restrictive, (ii) Exact prior membership probabilities may not be available for non-uniform sampling mechanisms such as snowball sampling. We provide methodological extensions to address these limitations in Appendix A.8.

5 EXPERIMENTS

Our experiments ask four questions: **RQ1:** Which factors drive membership privacy risk in non-private GNNs? **RQ2:** How does the privacy assessment change when using BMP-R instead of MP? **RQ3:** What are the practical implications of protecting membership and non-membership asymmetrically? **RQ4:** How sensitive are BMP estimates to the amount of auditing evidence on a larger graph? We study RQ1–RQ3 on CORA and CITESEER using GCN (Kipf & Welling, 2017) and GAT (Veličković et al., 2017), and RQ4 on the larger REDDIT graph using GRAPHSAGE (Hamilton et al., 2017), varying both the number of shadow models and target nodes. Implementation and MCMC details appear in Appendix A.7; datasets, models, and hyperparameters appear in Appendix A.9.

5.1 RESULTS

To answer RQ1-RQ3, we use CORA and CITESEER and vary GNN architecture (GCN vs GAT), training fraction (25% vs 50%), and training-graph sampling (uniform random with snowball sampling). Snowball sampling starts from a random seed and repeatedly adds randomly selected neighbors until the target fraction is reached, producing a connected training subgraph.

Attack-related variants. We run Algorithm 1 with $J = 1000$ shadow models trained using the target algorithm $\mathcal{A}$. For snowball sampling, we estimate each node's membership prior by its frequency in the shadow training sets and treat this estimate as fixed when inferring BMP parameters. We compare three pairs of attack variants: *strong vs. weak*, where the strong test uses the prior-aware likelihood ratio in Eq. (4) with $\tau = 1$ and the weak test sets $\eta_\nu = 1$; *clean vs. noisy*, where the noisy variant adds noise to the trained model's final layer; and *full vs. no edges*, where edges are retained or removed when querying the model. The base training procedure is the same across these variants. Beyond those comparisons, **we audit the node-level differentially private model GAP** (Sajadmanesh et al., 2023) and compare its privacy behavior with that of the non-private GCN and GAT models in Appendix A.10.3.

Privacy risk drivers in non-private GNNs. Tables 2, 3, 4, 5 in Appendix A.10 show the 5th, 50th (median), and 95th percentile (p5 / p50 / p95) estimates of the privacy parameters for MP and BMP-R with CORA with 25% and 50% sampling rates and CITESEER with 25% and 50% sampling rates, respectively. Figure 2 summarizes the results in those tables in terms of the marginal effect $\Delta = \hat{\varepsilon}(A) - \hat{\varepsilon}(B)$ of each experimental design choice on the posterior mean estimates $\hat{\varepsilon}$ of privacy parameters under choices A and B. Positive values indicate A is less private than B. The boxplots are obtained from the differences across all experimental configurations not involving the factor under consideration, thereby isolating the marginal contribution of that factor.

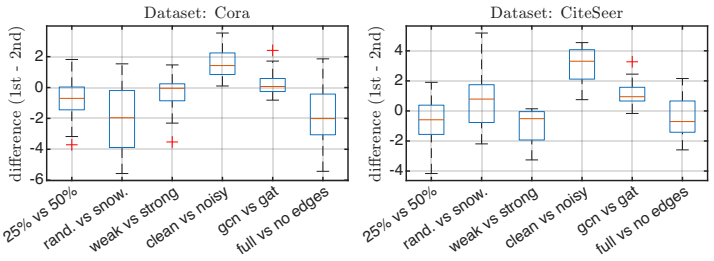


Figure 2: Marginal effects of experiment parameters on ϵ_R-BMP-R.

Figure 2 shows that training fraction (25% vs. 50%) and noise injection (clean vs. noisy scores) have the largest effects on ε_R-BMP-R for both CORA and CITESEER. We expected snowball sampling to reduce privacy by selecting nodes through graph connections, as observed on CORA. On the sparser CITESEER graph, however, snowball sampling appears slightly more private: its lower density weakens the structural signal created by following edges during sampling. Table 2 further shows that edge availability and GNN architecture have larger effects on CORA under random sampling.

Comparing MP and BMP-R. The detailed results in Tables 2-5 also reveal that in almost every configuration (except under random sampling with a 50% rate), the BMP-R privacy parameter estimates are significantly smaller than those of MP. This suggests that nodes with high likelihood ratios (suggesting high ε_{MP}) can have small membership priors, dampening their posterior membership probability and yielding $\varepsilon_R < \varepsilon_{\text{MP}}$. Figure 3 illustrates this relationship geometrically for a specific experiment: (CORA, 25% random sampling, GCN, clean outputs, full edges, strong attack). The largest per-node MP bound exceeds the largest BMP-R bound, with several high-MP-bound nodes having low prior odds.

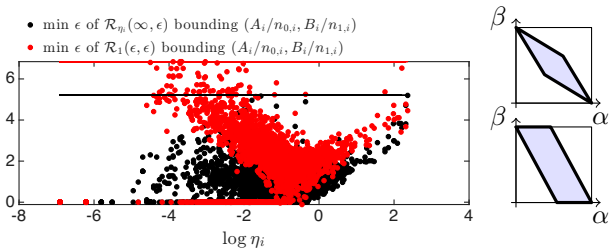


Figure 3: Min. per-node BMP-R (black) vs. MP (red) bounds for the empirical error rates. Horizontal lines mark the largest bounds. Right: Ex. MP and BMP-R regions.

Practical implications of asymmetry in BMP. Figure 4 compares joint estimation of general $(\varepsilon_L, \varepsilon_R)$-BMP for random and snowball sampling for CORA for a 50% split size. Snowball sampling shifts the posterior mass toward substantially larger values of both privacy parameters, indicating higher privacy leakage than random sampling. This exemplifies the effect of the sampling mechanism in BMP. We also observe that, in snowball sampling, the posterior estimates of ε_L and ε_R significantly differ, suggesting different upper bounds for the posterior membership and non-membership probabilities, thereby justifying the two separate parameters of BMP.

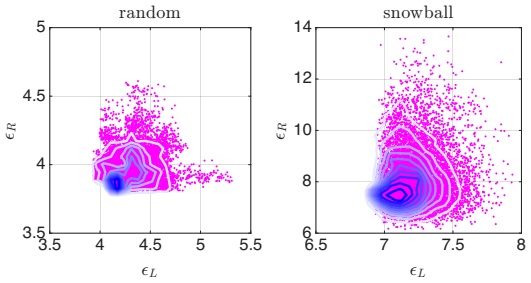


Figure 4: Post. samples of $(\varepsilon_R, \varepsilon_L)$ of BMP on CORA (50% samp., GCN, clean outputs).

5.1.1 SENSITIVITY TO AMOUNT OF AUDITING EVIDENCE

We use REDDIT to assess how BMP attack outcomes and inferred privacy parameters vary with the number of shadow models and audited nodes. Both quantities affect audit cost, allowing us to examine the sensitivity of privacy estimates to the computational budget. Each GraphSAGE shadow model is trained on 25% of Reddit's official training split, selected using snowball sampling. We use *clean outputs, full-edge inference, and the strong test* throughout.

Effect of the number of shadow models. We vary the number of shadow models over $J \in \{400, 600, 800, 1000, 2000, 3000\}$. Figure 7 (Appendix A.10.4) shows that the empirical membership priors and type-I/II error rates stabilize as J increases, with decreasing dispersion around their $J = 3000$ values. Table 1 reports the corresponding BMP posterior p5/p50/p95, fixing membership

priors to their estimates at $J = 3000$. The interval for ϵ_R progressively tightens with J, while that for ϵ_L is already narrow at $J = 400$, indicating that ϵ_L is largely determined by the prior-induced boundary in this setting. Full posterior distributions are provided in Figure 8 in Appendix A.10.4.

Table 1: BMP posterior p5/p50/p95 by the number of shadow-models J.

	$J = 400$	$J = 600$	$J = 800$	$J = 1000$	$J = 2000$	$J = 3000$
ε_L	7.31/7.33/7.36	7.31/7.32/7.36	7.32/7.33/7.37	7.31/7.32/7.35	7.31/7.32/7.36	7.31/7.32/7.35
ε_R	7.33/7.54/8.80	7.33/7.54/8.78	7.33/7.53/8.71	7.32/7.51/8.56	7.32/7.49/8.42	7.33/7.48/8.48

Effect of the number of audited nodes We study how node-level evidence affects the privacy audit while keeping the learning mechanism and shadow-model pool fixed. Starting from $K = 4,988$ target nodes evaluated with $J = 1000$ shadow models, we construct 10 nested subsets containing $10\%, 20\%, \ldots, 100\%$ of the nodes. For each subset, we estimate MP and jointly estimate $(\varepsilon_L, \varepsilon_R)$ for BMP, with node membership priors fixed to their empirical estimates.

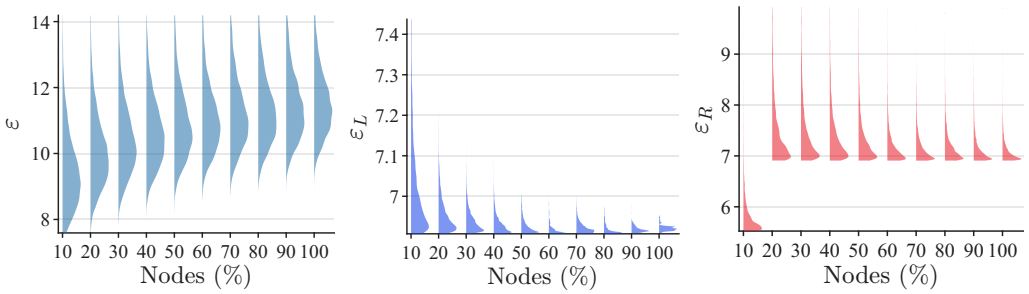


Figure 5: Posterior distribution vs node rate for ϵ-MP (left) and (ϵ_L, ϵ_R)-BMP (middle and right).

Figure 5 shows the resulting marginal distributions. The observed effect of adding audited nodes depends on the privacy definition. For MP, adding more nodes shifts the mean estimate of ϵ upwards until around 70% node rate; this suggests that inclusion of new nodes with 'extreme' error rates continues until that rate. From 70% on, the posterior stabilizes around $\epsilon \approx 11$. The shift in the posterior mean with earlier added nodes is also observed in BMP (especially for ϵ_R); however, stabilization of the overall posterior distribution occurs earlier than in MP. We also observe in the BMP case that the posterior distribution gets concentrated with increasing node rate.

6 Discussion and Conclusion

We introduced BMP, a sampling-aware framework for analyzing node-level membership leakage in GNNs within a Bayesian framework. By incorporating node-dependent priors and casting membership inference as Bayesian hypothesis testing, BMP captures heterogeneous privacy risk that uniform metrics may miss. Our designed MIA and the accompanying MCMC posterior sampling method enable computation of uncertainty-aware, attack-dependent, model-based estimates of privacy parameters for arbitrary GNN mechanisms.

The auditing method in Section 4 is based on Theorem 4, which characterizes BMP through the error probabilities of all decision rules and nodes. However, a practical audit evaluates only a finite set of target nodes using a particular attack. Consequently, the observed attack performance provides attack-dependent empirical evidence and may yield a lower bound on worst-case privacy loss; it cannot rule out stronger attacks. Our Bayesian procedure supplements this evidence with the probabilistic modeling assumptions in Section 4.2 to obtain uncertainty-aware, model-based estimates. Such numerical auditing complements analytical guarantees, which are often unavailable or loose for complex learning mechanisms such as GNNs. We follow the growing literature on empirical privacy evaluation and estimation using MIAs (Yeom et al., 2018; Carlini et al., 2022; Jayaraman & Evans, 2019; Humphries et al., 2023; Zanella-Beguelin et al., 2023; Yıldırım et al., 2026), extending this approach to graph sampling and GNNs.

AI USE STATEMENT

In this work, we used generative AI tools to restructure text and correct mathematical and grammatical mistakes. We benefited from the conference's PAT automated pre-submission feedback. We also used AI tools to assist in coding. Research ideation, methodological development and decisions, theoretical and experimental analyses, and conclusions were all made by the authors.

ETHICS STATEMENT

This work studies node-level membership privacy in graph neural networks and proposes a Bayesian framework for auditing privacy leakage from trained models. The empirical evaluation uses the established benchmark graph datasets CORA, CITESEER and REDDIT. Those are available benchmark datasets, and we will verify and comply with their applicable terms of use and citation requirements. We do not conduct human-subject experiments, collect new personal or sensitive information, or attempt to identify individuals represented in any dataset.

The work is motivated by the privacy risks of graph-based machine learning systems. In particular, membership inference may reveal whether a node was included in a training graph, which could be sensitive in real-world applications. Our experiments are designed to quantify such risks under different sampling strategies and model configurations. The examples involving potentially sensitive applications are illustrative and are not based on real-world user data.

REPRODUCIBILITY STATEMENT

We release our code anonymously at `https://anonymous.4open.science/r/mia_project-8963/README.md`.

The Appendix is organized as follows:

- Appendix A.1 contains Algorithm 1, which measures the performance of the MIA described in Section 4.1, and was used in our experiments to produce the error counts.
- Appendix A.2 contains three example cases where membership privacy is justified.
- Appendix A.3 contains a review of related work that positions the current work.
- Appendix A.4 contains differential privacy (DP)-variants of the membership privacy definitions as a complementary of Section 3.
- Appendix A.5 contains the missing proofs of the theoretical results in Section 3.1.
- Appendix A.6 has Algorithm 2, the MCMC algorithm that was used to estimate the privacy parameters.
- Appendix A.7 contains the implementation details for Bayesian estimation.
- Appendix A.8 discusses the limitations of the modeling and methodological approaches, and provides concrete suggestions to address them.
- Appendix A.9 provides detailed information on the datasets and the details of the GNN models implemented in Section 5.
- Section A.10 presents the detailed experimental results.

REFERENCES

Christophe Andrieu, Sinan Yıldırım, Arnaud Doucet, and Nicolas Chopin. Metropolis-hastings with averaged acceptance ratios, 2020. URL `https://arxiv.org/abs/2101.01253`.

Jiamu Bai, Guanlin He, Xin Gu, Daniel Kifer, and Kiwan Maeng. Composition for pufferfish privacy, 2026. URL `https://arxiv.org/abs/2602.02718`.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914. IEEE, 2022.

Enyan Dai, Tianxiang Zhao, Huaisheng Zhu, Junjie Xu, Zhimeng Guo, Hui Liu, Jiliang Tang, and Suhang Wang. A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability. *arXiv preprint arXiv:2204.08570*, 2022.

Ameya Daigavane, Gagan Madan, Aditya Sinha, Abhradeep Guha Thakurta, Gaurav Aggarwal, and Prateek Jain. Node-level differentially private graph neural networks. In *ICLR 2022 Workshop on PAIR^2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*, 2022. URL https://openreview.net/forum?id=BCfgOLx3gb9.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

Longzhu He, Chaozhuo Li, Peng Tang, and Sen Su. Going deeper into locally differentially private graph neural networks. In *Forty-second International Conference on Machine Learning*, 2025a. URL https://openreview.net/forum?id=2aKHuXdr7Q.

Longzhu He, Chaozhuo Li, Peng Tang, and Sen Su. Going deeper into locally differentially private graph neural networks. In *Forty-second International Conference on Machine Learning*, 2025b.

Xinlei He, Jinyuan Jia, Michael Backes, Neil Zhenqiang Gong, and Yang Zhang. Stealing links from graph neural networks. In *30th USENIX security symposium (USENIX security 21)*, pp. 2669–2686, 2021.

Thomas Humphries, Simon Oya, Lindsey Tulloch, Matthew Rafuse, Ian Goldberg, Urs Hengartner, and Florian Kerschbaum. Investigating membership inference attacks under data dependencies. In *2023 IEEE 36th Computer Security Foundations Symposium (CSF)*, pp. 473–488. IEEE, 2023.

Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: how private is private sgd? In *Proc. of the 34th Int. Conf. on Neural Information Proc. Sys.*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *28th USENIX security symposium (USENIX security 19)*, pp. 1895–1912, 2019.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. *IEEE Trans. Inf. Theor.*, 63(6):4037–4049, June 2017. ISSN 0018-9448.

Megha Khosla. Impact of graph structure on membership-inference risk for graph neural networks. *Proceedings of Privacy Enhancing Technologies Symposium (PETS 2026)*, 2026.

Daniel Kifer and Ashwin Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Trans. Database Syst.*, 39(1), January 2014. ISSN 0362-5915. doi: 10.1145/2514689. URL https://doi.org/10.1145/2514689.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

Jaewoo Lee and Chris Clifton. Differential identifiability. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pp. 1041–1049, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450314626. doi: 10.1145/2339530.2339695. URL https://doi.org/10.1145/2339530.2339695.

Dian Lei, Zijun Song, Yanli Yuan, Chunhai Li, and Liehuang Zhu. Achieving personalized privacy-preserving graph neural network via topology awareness. In *THE WEB CONFERENCE 2025*, 2025. URL https://openreview.net/forum?id=hGdbYi3hnB.

Ninghui Li, Wahbeh Qardaji, Dong Su, Yi Wu, and Weining Yang. Membership privacy: A unifying framework for privacy definitions. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pp. 889–900, 2013.

Milad Nasr, Shuang Songi, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlin. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on security and privacy (SP)*, pp. 866–882. IEEE, 2021.

Iyiola E Olatunji, Wolfgang Nejdl, and Megha Khosla. Membership inference attack on graph neural networks. In *2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, pp. 11–20. IEEE, 2021.

Iyiola E Olatunji, Thorben Funke, and Megha Khosla. Releasing graph neural networks with differential privacy guarantees. *Transactions on Machine Learning Research*, 2023.

Sina Sajadmanesh and Daniel Gatica-Perez. Progap: Progressive graph neural networks with differential privacy guarantees. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, pp. 596–605, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703713. doi: 10.1145/3616855.3635761. URL https://doi.org/10.1145/3616855.3635761.

Sina Sajadmanesh, Ali Shahin Shamsabadi, Aurélien Bellet, and Daniel Gatica-Perez. Gap: differentially private graph neural networks with aggregation perturbation. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23, USA, 2023. USENIX Association. ISBN 978-1-939133-37-3.

Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*, 2018.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Aleksei Triastcyn and Boi Faltings. Bayesian differential privacy for machine learning. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Xiuling Wang and Wendy Hui Wang. Subgraph structure membership inference attacks against graph neural networks. *Proceedings on Privacy Enhancing Technologies*, 2024.

Yu Wang, Lifu Huang, Philip S Yu, and Lichao Sun. Membership inference attacks on knowledge graphs. *arXiv preprint arXiv:2104.08273*, 2021.

Bang Wu, Xiangwen Yang, Shirui Pan, and Xingliang Yuan. Adapting membership inference attacks to gnn for graph classification: Approaches and implications. In *2021 IEEE International Conference on Data Mining (ICDM)*, pp. 1421–1426. IEEE, 2021.

Zihang Xiang, Tianhao Wang, and Di Wang. Preserving node-level privacy in graph neural networks. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 4714–4732. IEEE, 2024.

Hanshen Xiao and Srinivas Devadas. Pac privacy: Automatic privacy measurement and control of data processing. In Helena Handschuh and Anna Lysyanskaya (eds.), *Advances in Cryptology – CRYPTO 2023*, pp. 611–644, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-38545-2.

Bin Yang, Issei Sato, and Hiroshi Nakagawa. Bayesian differential privacy on correlated data. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, pp. 747–762, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450327589. doi: 10.1145/2723372.2747643. URL https://doi.org/10.1145/2723372.2747643.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pp. 268–282. IEEE, 2018.

Ceren Yıldırım, Kamer Kaya, Sinan Yıldırım, and Erkay Savaş. Mcmc for bayesian estimation of differential privacy from membership inference attacks. In Rita P. Ribeiro, Bernhard Pfahringer, Nathalie Japkowicz, Pedro Larrañaga, Alípio M. Jorge, Carlos Soares, Pedro H. Abreu, and João Gama (eds.), *Machine Learning and Knowledge Discovery in Databases. Research Track*, pp. 397–414, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-06096-9.

Santiago Zanella-Beguelin, Lukas Wutschitz, Shruti Tople, Ahmed Salem, Victor Rühle, Andrew Paverd, Mohammad Naseri, Boris Köpf, and Daniel Jones. Bayesian estimation of differential privacy. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 40624–40636. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/zanella-beguelin23a.html.

Qiuchen Zhang, Hong kyu Lee, Jing Ma, Jian Lou, Carl Yang, and Li Xiong. Dpar: Decoupled graph neural networks with node-level differential privacy. In *Proceedings of the ACM Web Conference 2024*, WWW '24, pp. 1170–1181, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701719. doi: 10.1145/3589334.3645531. URL https://doi.org/10.1145/3589334.3645531.

A Appendix

A.1 Measuring Protocol for Node-MIA

Algorithm 1 Measuring Protocol for Node-MIA

Require: Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, algorithm $\mathcal{A} = (\mathcal{P}, \Phi)$, target set $\mathcal{T} = \{\nu_1, \ldots, \nu_K\}$, number of shadow models J, threshold $\tau > 0$

Ensure: True counts $n_{0,1:K}$, $n_{1,1:K}$; errors $A_{1:K}$, $B_{1:K}$

for $j = 1, \ldots, J$ **do**
  $S_j \sim \mathcal{P}(\mathcal{G})$, $\phi_j = \Phi(\mathcal{G}_{S_j}, \mathbf{X}_{S_j}, \mathbf{Y}_{S_j})$. ▷ Sample J training graph and train GNN on each
for $k = 1, \ldots, K$ **do**
  $n_{0,k} = \sum_{j=1}^{J} \mathbb{I}(\nu_k \notin S_j)$, $n_{1,k} = J - n_{0,k}$, $\hat{\eta}_k = n_{1,k}/n_{0,k}$ ▷ True counts
  Set $A_k = 0$, $B_k = 0$,
  for $j \in \{1, \ldots, J\}$ **do** ▷ Test every shadow model for the membership of ν_k
    Fit $\mathcal{N}(\mu_0, \sigma_0^2)$ to $\mathcal{Z}_0 = \{z_\nu(\phi_i) : i \in \{1, \ldots, J\} \setminus \{j\}, \nu \notin S_i\}$ ▷ Learn H_0
    Fit $\mathcal{N}(\mu_1, \sigma_1^2)$ to $\mathcal{Z}_1 = \{z_\nu(\phi_i) : i \in \{1, \ldots, J\} \setminus \{j\}, \nu \in S_i\}$ ▷ Learn H_1
    Calculate $\hat{m} = \hat{m}_\tau(z_{\nu_k}(\phi_j))$ using $\mathcal{N}(\mu_0, \sigma_0^2)$, $\mathcal{N}(\mu_1, \sigma_1^2)$, $\hat{\eta}_k$, and τ. ▷ Decision
    $A_k \leftarrow A_k + \mathbb{I}(\hat{m} = 1, \nu_k \notin S_j)$ ▷ Check type-I error
    $B_k \leftarrow B_k + \mathbb{I}(\hat{m} = 0, \nu_k \in S_j)$ ▷ Check type-II error

A.2 Example cases for Membership privacy

Example 1 ($\mathcal{P} \neq \mathcal{P}_*$*, membership is sensitive due to hidden sensitive selection criteria*)**.** Consider a social graph such as Facebook's. Suppose a platform trains a model on a subset of users selected for a mental-health intervention classifier. The training will use only publicly observable features $(\mathbf{X}, \mathbf{Y})$ such as friendship graph, age range, etc, which are known to the adversary. However, the training subset deliberately consists of users internally flagged as likely vulnerable based on private signals ($\mathbf{Z}$) unavailable publicly (e.g., help-center interactions, deleted posts). Hence, $\mathcal{P} \neq \mathcal{P}_*$. Then learning that the user i belonged to the training subset reveals that user i was considered part of a sensitive target population, which is a privacy breach of the individual.

Example 2 ($\mathcal{P} \neq \mathcal{P}_*$*, non-membership is sensitive due to hidden sensitive selection criteria*)**.** A company trains an internal model on applicants shortlisted for executive leadership potential. The shortlist ($\mathbf{Z}$) is confidential; from the adversary's view, the sampling is random among the employees; hence $\mathcal{P} \neq \mathcal{P}_*$. Then, inferring that an employee is a non-member may reveal that they were not considered for promotion, which can be highly sensitive professionally.

Example 3 ($\mathcal{P} = \mathcal{P}_*$*, membership is sensitive due to sensitive post-training treatment*)**.** A medical biobank trains a GNN on a subset of consenting participants using existing measurements and family relationships. The selection is truly random and publicly known; hence, $\mathcal{P} = \mathcal{P}_*$. Under a publicly known study protocol, all selected participants subsequently undergo whole-genome sequencing, but their identities remain confidential. Inferring membership therefore reveals whose genome the institution sequences—information that participants may wish to keep private.

A.3 Relation to Bayesian privacy definitions

Our privacy definitions are closest to the posterior membership privacy PMP definitions of Li et al. (2013). In both our work and Li et al. (2013), the population (the whole dataset) is assumed to be known, and the sensitive unknown information is the membership of the elements in the population. Also, both our definitions consider the posterior membership probabilities, and we note that some one-to-one conversions are possible in special cases. However, there are a few differences between the approaches: While PMP focuses on the difference between the posterior and the prior as a metric of informativeness of the algorithm, BMP focuses directly on the posterior, reflecting our view that the sampling phase is an integral part of the algorithm (hence the notation $\mathcal{A} = (\mathcal{P}, \Phi)$). Also, while Li et al. (2013) introduce "positive PMP" and "negative PMP" (in parallel to BMP-R and BMP-L) corresponding to membership and non-membership posterior probabilities, our BMP Definition 2 unifies the two cases. Finally, while we consider a single sampling distribution $\mathcal{P}$ attached to the algorithm $\mathcal{A}$, Li et al. (2013) assumes a set of sampling distributions for which the defining inequality holds.

Differential identifiability of Lee & Clifton (2012) is also related to our work in that their definition corresponds to the ε-BMDP-R. There also exist related definitions that employ the Bayesian paradigm, designed for general data structures, yet applicable to membership information as a special case. For example, Yang et al. (2015) propose a Bayesian DP (BDP) definition for correlated sensitive data. BDP concerns the likelihood ratio involving an unknown component of the sensitive data, conditional on a known part of the sensitive data. The MP in Definition 1 corresponds to the BDP of Yang et al. (2015) when the membership information is viewed as correlated data, so that the unknown element is a node, and no membership information is assumed known. Unlike our approach, the term "Bayesian" is used not because of the involvement of the posterior distribution of the unknown element in question; it is rather because of assigning a probability distribution on the sensitive data so that marginal likelihoods can be computed via marginalization. Another definition, with the same name BDP, is proposed in Triastcyn & Faltings (2020). The definition is reminiscent of the classical DP, with the difference that the ratio of likelihoods is treated as a random variable by employing the prior of the differing element between the datasets.

Our definition also bears similarities to PAC Membership Privacy (Xiao & Devadas, 2023, Def. 5): both are non-differential notions and assume that the adversary knows the population and the sampling process, but not the realized sampled training set. A key distinction, however, is that BMP explicitly supports non-uniform, node-dependent membership priors induced by graph sampling, making it better suited to graph settings with heterogeneous inclusion probabilities.

A.4 DP versions of membership privacy definitions

We define conditional membership probabilities. For any vector $m_{1:N} \in \{0,1\}^N$, let

$$\gamma_i^c(m_i \mid m_{-i}) := \Pr(M_i = m_i \mid M_{-i} = m_{-i}).$$

where $M_{-i} := (M_{1:i-1}, M_{i+1:N})$ (similarly for m_{-i}).

We also define conditional posterior probabilities conditional on $\mathcal{A}$'s output. For any $i \in \{1, \ldots, N\}$ and $m_{1:N} \in \{0,1\}^N$, and $w \in \mathcal{W}$ such that $f_0(w) > 0$, define the full conditional posteriors

$$\pi_i^c(m_i | m_{-i}, w) := \Pr(M_i = m_i | M_{-i} = m_{-i}, W = w).$$

Definition 5 (Membership DP (MDP)). *We say $\mathcal{A}$ is ε-MDP if for any $w \in \mathcal{W}$ such that $f_0(w) > 0$, and for any $i \in \{1, \ldots, N\}$ and $m_{-i} \in \{0,1\}^{N-1}$, we have*

$$e^{-\varepsilon} \leq \frac{f(w|m_{1:i-1}, 1, m_{i+1:N})}{f(w|m_{1:i-1}, 0, m_{i+1:N})} \leq e^{\varepsilon}.$$

Definition 6 (Bayesian MDP (BMDP)). *$\mathcal{A} = (\mathcal{P}, M)$ is $(\varepsilon_L, \varepsilon_R)$-BMDP if for every $i \in \{1, \ldots, N\}$, $m_{-i} \in \{0,1\}^{N-1}$, output $w \in \mathcal{W}$ with $f_0(w) > 0$,*

$$e^{-\varepsilon_L} \leq \pi_i^c(1|m_{-i}, w)/\pi_i^c(0|m_{-i}, w) \leq e^{\varepsilon_R} \tag{8}$$

The special cases of ε-BMDP-L and ε-BMDP-R can also be defined similarly as (ε, ∞)-BMDP and (∞, ε)-BMDP. In fact, ε-BMDP-R corresponds to the differential identifiability of Lee & Clifton (2012).

A.5 MISSING PROOFS

Proof of Proposition 1. We will only prove the first inequality; the second one follows similarly. Fix $i \in \{1, \ldots, N\}$. We have $\mathbb{E}(\pi_i(1|W)) = \gamma_i$, where the expectation is with respect to the marginal distribution W (with density $f_0(\cdot)$). This means that there is value $w \in \mathcal{W}$ with $f_0(w) > 0$ such that $\pi_i(1|w) \geq \gamma_i(1)$, which implies that $\pi_i(1|w)/\pi_i(0|w) \geq \gamma_i(1)/\gamma_i(0)$. Since $\mathcal{A}$ is ε-BMP, we have $e^{\varepsilon_R} \geq \pi_i(1|w)/\pi_i(0|w) \geq \gamma_i(1)/\gamma_i(0)$. Since the above holds for any $i \in \{1, \ldots, N\}$, we have the claimed lower bound for ε_R. □

For the proofs of Proposition 2 and Theorems 4 and 5, note the useful fact that the definitions in Section 3 can be written in terms of output sets with positive probability, i.e.,

$$e^{-\varepsilon_L} \leq \frac{\pi_i(1|w)}{\pi_i(0|w)} \leq e^{\varepsilon_R} \Leftrightarrow e^{-\varepsilon_L} \leq \frac{\pi_i(1|O)}{\pi_i(0|O)} \leq e^{\varepsilon_R}$$

for any $O \subseteq \mathcal{W}$ with $\Pr(W \in O) > 0$ the marginal and full conditional posteriors

$$\pi_i(m_i|O) := \Pr(M_i = m_i | W \in O).$$

Proof of Proposition 2. To distinguish between the posterior membership probabilities conditional on the outputs of $\mathcal{A}$ and $\mathcal{B}$, we denote those probabilities as $\pi^{\mathcal{A}}$ and $\pi^{\mathcal{B}}$, respectively. Let $W = \Phi(\mathcal{G}_S, X_S)$ with $S \sim \mathcal{P}(\mathcal{G})$ and let $Z = \varphi(\Phi(\mathcal{G}_S, X_S))$ be the output of $\mathcal{B}$. Fix $i \in \{1, \ldots, N\}$. For any $O_z \subseteq \mathcal{Z}$, we have $\{Z \in O_z\} = \{W \in O_y\}$, where $O_w = \varphi^{-1}(O_z) = \{w \in \mathcal{W} : \varphi(w) \in O_z\} \subseteq \mathcal{W}$. Hence, $\pi_i^{\mathcal{B}}(1|O_z) := \Pr(M_i = 1|Z \in O_z) = \Pr(M_i = 1|W \in O_w) = \pi_i^{\mathcal{A}}(1|O_w)$. Since $\mathcal{A}$ is $(\varepsilon_L, \varepsilon_R)$,

$$e^{-\varepsilon_L} \leq \pi_i^{\mathcal{A}}(1|O_w)/\pi_i^{\mathcal{A}}(0|O_w) = \pi_i^{\mathcal{B}}(1|O_z)/\pi_i^{\mathcal{B}}(0|O_z) \leq e^{\varepsilon_R}.$$

Hence, the claim is proven. □

Proof of Proposition 3. Assume $\mathcal{A}$ is ε-MP. Then for any $w \in \mathcal{W}$, and for all $i \in \{1, \ldots, N\}$, we have

$$e^{-\varepsilon}\eta_i \leq \frac{\pi_i(1|w)}{\pi_i(0|w)} \leq \eta_i e^{\varepsilon}$$

or

$$e^{-(\varepsilon - \log \eta_i)} \leq \frac{\pi_i(1|w)}{\pi_i(0|w)} \leq e^{\varepsilon + \log \eta_i}$$

A uniform bound can be developed for the ratio $\frac{\pi_i(1|w)}{\pi_i(0|w)}$ as

$$e^{-(\varepsilon - \log \eta_{\min})} \leq \frac{\pi_i(1|w)}{\pi_i(0|w)} \leq e^{\varepsilon + \log \eta_{\max}}$$

Therefore, $\mathcal{A}$ is $(\varepsilon_L, \varepsilon_R)$-BMP, with

$$\varepsilon_L = \varepsilon - \log \eta_{\min}, \quad \varepsilon_R = \varepsilon + \log \eta_{\max}.$$

For the other direction, assume $\mathcal{A}$ is $(\varepsilon_L, \varepsilon_R)$-BMP. Then

$$e^{-\varepsilon_L} \leq \frac{\pi_i(1|w)}{\pi_i(0|w)} \leq e^{\varepsilon_R},$$

or,

$$e^{-(\varepsilon_L + \log \eta_i)} \leq \frac{f_i(w|1)}{f_i(w|0)} \leq e^{\varepsilon_R - \log \eta_i},$$

A uniform bound can be developed for the ratio $\frac{f_i(w|1)}{f_i(w|0)}$ as

$$e^{-(\varepsilon_L + \log \eta_{\max})} \leq \frac{f_i(w|1)}{f_i(w|0)} \leq e^{\varepsilon_R - \log \eta_{\min}}$$

Considering the maximum of the two exponents, we conclude that $\mathcal{A}$ is ε-MP where

$$\varepsilon = \max\left\{\varepsilon_L + \log \eta_{\max}, \varepsilon_R - \log \eta_{\min}\right\}.$$

□

Proof of Theorem 4. For any $i \in \{1, \ldots, N\}$, $m_i \in \{0, 1\}$, and $O \subseteq \mathcal{W}$, let $F_i(O|m_i) := \Pr(W \in O|M_i = m_i)$. Any decision rule based on the output of $\mathcal{A}$ can be written in terms of a critical region $C \subseteq \mathcal{W}$ as $D(W) = \mathbb{I}(W \in C)$ for a measurable set C. Definition 2 guarantees that $\gamma_i(0)F_i(C|0) \leq e^{\varepsilon_L}\gamma_i(1)F_i(C|1)$. Using $F_i(C|0) = \alpha_i$, $1 - F_i(C|1) = \beta_i$, we have

$$\eta_i^{-1}e^{-\varepsilon_L}\alpha_i + \beta_i \leq 1$$

Definition 2 also implies $\gamma_i(1)F_i(C|1) \leq e^{\varepsilon_R}\gamma_i(0)F_i(C|0)$. Using the same substitutions for α_i and β_i, we get

$$\eta_i^{-1}e^{\varepsilon_R}\alpha_i + \beta_i \geq 1.$$

Next, applying Definition 2 with the critical region being C^c so that $F_i(C^c|0) = 1-\alpha_i$, $F_i(C^c|1) = \beta_i$. We have $\gamma_i(0)F_i(C^c|0) \leq e^{\varepsilon_L}\gamma_i(1)F_i(C^c|1)$. Using the substitutions for α_i and β_i, we get

$$\alpha_i + \eta_i e^{\varepsilon_L}\beta_i \geq 1$$

Finally, the right-hand inequality for C^c reads $\gamma_i(1)F_i(C^c|1) \leq e^{\varepsilon_R}\gamma_i(0)F_i(C^c|0)$, which yields

$$\eta_i e^{-\varepsilon_R}\beta_i + \alpha_i \leq 1.$$

Hence, all the inequalities in $\mathcal{R}_i(\varepsilon_L, \varepsilon_R)$ are satisfied. For the reverse implication, it suffices to go backwards from each equality to arrive at the inequality in the Definition 2. □

Proof of Theorem 5. For a given decision rule D, the expected cost is

$$\begin{aligned}
\mathbb{E}[\text{Cost}] =& c_0 \int \mathbb{I}_1(D(w))\gamma_i(0)f_i(w|0)\mathrm{d}w \\
&+ c_1 \int \mathbb{I}_0(D(w))\gamma_i(1)f_i(w|1)\mathrm{d}w \\
=& \int [c_0\pi_i(0|w)\mathbb{I}_1(D(w)) + c_1\pi_i(1|w)\mathbb{I}_0(D(w))]f_0(\mathrm{d}w) \\
=& \mathbb{E}[c_0\pi_i(0|W)\mathbb{I}_1(D(W)) + c_1\pi_i(1|W)\mathbb{I}_0(D(W))]
\end{aligned}$$

This is minimized when

$$D(w) = \begin{cases} 1 & c_0\pi_i(0|w) \leq c_1\pi_i(1|w) \\ 0 & c_0\pi_i(0|w) > c_1\pi_i(1|w) \end{cases}, \quad w \in \mathcal{W}$$

and the minimum expected cost E_{min} is

$$E_{min} = \mathbb{E}[\min\{c_0\pi_i(0|W), c_1\pi_i(1|W)\}].$$

By Definition 2, $e^{-\varepsilon_L} \leq \pi_i(1|W)/\pi_i(0|W) \leq e^{\varepsilon_R}$, which yields

$$\min\{c_0\pi_i(0|W), c_1\pi_i(1|W)\} \geq \min\left\{\frac{c_0}{1+e^{\varepsilon_R}}, \frac{c_1}{1+e^{\varepsilon_L}}\right\}$$

On the other hand, since $\pi_i(0|W)/\pi_i(1|W) \geq e^{-\varepsilon_R}$, we have

$$\begin{aligned}
E_{\min} &= \mathbb{E}\left[\pi_i(1|W)\min\left\{c_0\frac{\pi_i(0|W)}{\pi_i(1|W)}, c_1\right\}\right] \\
&\geq \mathbb{E}\left[\pi_i(1|W)\min\{c_0e^{-\varepsilon_R}, c_1\}\right] \\
&\geq \gamma_i(1)\min\{c_0e^{-\varepsilon_R}, c_1\}
\end{aligned}$$

where in the last line we have used the fact that $\mathbb{E}[\pi_i(1|W)] = \gamma_i(1)$. By following similar steps but using $\pi_i(1|W)/\pi_i(0|W) \geq e^{-\varepsilon_L}$, we also have

$$E_{\min} \geq \gamma_i(0)\min\{c_1e^{-\varepsilon_L}, c_0\}$$

The three lower bounds can be combined to yield (3). In particular, the probability of a wrong decision is obtained by $\mathbb{E}(\text{Cost})$ with $c_1 = c_2 = 1$, yielding the stated lower bound for $\Pr(M_i \neq D)$. □

Composing BMP with BMP. Composing T BMP mechanisms involves T sampling and training runs. Here we define membership of a node as its inclusion in at least one of the runs. Proposition 6 provides a composition result corresponding to this definition.

Proposition 6 (Composition: BMP-BMP). *Assume that algorithm $\mathcal{A}_t = (\mathcal{P}_t, \Phi_t)$ is $(\varepsilon_L^{(t)}, \varepsilon_R^{(t)})$-BMP, for $t = 1, \ldots, T$, and the sampling operations are independent. Define the composition algorithm $\mathcal{A} : (\mathcal{P}, \Phi)$, where $\mathcal{P} = \bigotimes_{t=1}^{T} \mathcal{P}_t$ is the product law and*

$$\Phi(S_{1:T}) = (\Phi_1(\mathcal{G}_{S_1}, \mathbf{X}_{S_1}, \mathbf{Y}_{S_1}), \ldots, \Phi_T(\mathcal{G}_{S_T}, \mathbf{X}_{S_T}, \mathbf{Y}_{S_T}))$$

applies Φ_t on the sampled dataset according to $S_t \sim \mathcal{P}_t$. Then $\mathcal{A}$ is $(\bar{\varepsilon}_L, \bar{\varepsilon}_R)$ BMP, where

$$\bar{\varepsilon}_L = \sum_{t=1}^{T} \log \mu_t - \log\left[1 - \prod_{t=1}^{T} \mu_t\right] \quad \bar{\varepsilon}_R = \log\left[1 - \prod_{t=1}^{T}(1-\kappa_t)\right] - \sum_{t=1}^{T} \log(1-\kappa_t),$$

where $\mu_t = e^{\varepsilon_L^{(t)}}/(1+e^{\varepsilon_L^{(t)}})$ and $\kappa_t = e^{\varepsilon_R^{(t)}}/(1+e^{\varepsilon_R^{(t)}})$.

Proof of Proposition 6. Let the membership information of the tth sampling be $M^{(t)}$. Then, given $W_{1:T} = w_{1:T}$

$$\Pr\Big(\exists t \in (1, \ldots, T) : M_i^{(t)} = 1 \mid W_{t:T} = w_{1:T}\Big) \leq 1 - \prod_{t=1}^{T}(1-\kappa_t)$$

where $\kappa_t = \frac{e^{\varepsilon_R^{(t)}}}{1+e^{\varepsilon_R^{(t)}}}$. Also, we can prove that

$$\Pr\Big(\forall t \in (1, \ldots, T) : M_i^{(t)} = 0 \mid W_{t:T} = w_{1:T}\Big) \leq \prod_{t=1}^{T} \mu_t$$

where $\mu_t = \frac{e^{\varepsilon_L^{(t)}}}{1+e^{\varepsilon_L^{(t)}}}$. Combining both inequalities, we can show that the composition is $(\varepsilon_L, \varepsilon_R)$-BMP, where

$$\varepsilon_L = \sum_{t=1}^{T} \log \mu_t - \log\left[1 - \prod_{t=1}^{T} \mu_t\right], \quad \varepsilon_R = \log\left[1 - \prod_{t=1}^{T}(1-\kappa_t)\right] - \sum_{t=1}^{T} \log(1-\kappa_t).$$

So, the claim is proven. □

A.6 MCMC ALGORITHM FOR BAYESIAN ESTIMATION OF BMP PARAMETERS

The MCMC algorithm for Bayesian estimation of BMP parameters is given in Algorithm 2.

A.7 IMPLEMENTATION DETAILS FOR BAYESIAN ESTIMATION

In all the Bayesian estimation tasks, we relied on the hierarchical model in (7) and the MHAAR algorithm in Algorithm 2 designed for that hierarchical model, with necessary modifications to suit the methodology for the given privacy definition. Specifically,

- For MP, we took $\varepsilon_{\text{MP}} \sim \mathcal{N}(0, 10)$.
- For BMP-R, we fixed $\varepsilon_L = \infty$ and assign the prior $\varepsilon_R \sim \mathcal{N}(0, 10)$;
- For BMP, we took $\varepsilon_L, \varepsilon_R \overset{\text{iid}}{\sim} \mathcal{N}(0, 10)$.

For all tasks, the attack strength parameter is fixed to $s = 0$, i.e., the (α_k, β_k)-error probabilities of a MIA attack are assumed to be uniformly distributed over their allowed region $\mathcal{R}_{\eta_{\nu_k}}(\varepsilon_L, \varepsilon_R)$. The reported experiments use the independent-binomial special case of g_θ, i.e., $A_k \sim \text{Binom}(n_{0,k}, \alpha_k)$ and $B_k \sim \text{Binom}(n_{1,k}, \beta_k)$, independently, so the parameter θ is absent from this model. The dependence-aware model in (9) can be substituted when one wants uncertainty quantification that explicitly accounts for the leave-one-out reuse of shadow models.

We used independent normal random-walk proposals for ε_L and ε_R, whose variances are adapted during MCMC iterations to achieve an acceptance rate of 0.23. In all experiments, the number of auxiliary variables M for (α_k, β_k) for each k is taken $M = 50$. To sample auxiliary variables for

Algorithm 2 MCMC for posterior sampling for $(\varepsilon_L, \varepsilon_R, \theta)$

Require: Counts of non-membership and membership instances $n_{0,1:K}, n_{1,1:K}$; error counts $A_{1:K}$, $B_{1:K}$;
Ratio of membership/non-membership priors: $\eta_{1:K}$;
prior distribution for BMP parameters $p(\varepsilon_L, \varepsilon_R)$; prior distribution $p(\theta)$ on θ;
proposal distributions $q(\varepsilon'_L, \varepsilon'_R|\varepsilon_L, \varepsilon_R)$, $q(\theta'|\theta)$; proposal distribution for (α_k, β_k), $Q_k(\alpha, \beta)$, $k = 1, \ldots, K$;
M: number of auxiliary variables for each node $k \in \{1, \ldots, K\}$;
T: Number of MCMC iterations;
Initial values: $\varepsilon_L, \varepsilon_R, \theta$.

Ensure: MCMC samples: $\{\varepsilon_L^{(t)}, \varepsilon_R^{(t)}, \theta^{(t)}; t = 1, \ldots, T\}$

for $t = 1 : T$ **do**

Draw the proposal $(\varepsilon'_L, \varepsilon'_R) \sim q(\varepsilon'_L, \varepsilon'_R|\varepsilon_L, \varepsilon_R)$, and $\theta' \sim q(\theta'|\theta)$.

for $k = 1 : K$ **do**

Set $(\alpha_k^{(1)}, \beta_k^{(1)}) = (\alpha_k, \beta_k)$; and sample $\alpha_k^{(i)}, \beta_k^{(i)} \overset{\text{iid}}{\sim} Q_k(\alpha, \beta)$ for $i = 2, \ldots, M$.

for $i = 1 : M$ **do**

Calculate the weights

$$w_k^{(i)} = \frac{h_{\eta_k}(\alpha_k^{(i)}, \beta_k^{(i)}|\varepsilon_L, \varepsilon_R) g_\theta(A_k, B_k|\alpha_k^{(i)}, \beta_k^{(i)})}{Q_k(\alpha_k^{(i)}, \beta_k^{(i)})},$$

$$w_k'^{(i)} = \frac{h_{\eta_k}(\alpha_k^{(i)}, \beta_k^{(i)}|\varepsilon'_L, \varepsilon'_R) g_{\theta'}(A_k, B_k|\alpha_k^{(i)}, \beta_k^{(i)})}{Q_k(\alpha_k^{(i)}, \beta_k^{(i)})}.$$

Decision: Accept the proposals $(\varepsilon_L, \varepsilon_R) = (\varepsilon'_L, \varepsilon'_R), \theta = \theta'$ and set $\bar{w}_{1:K}^{(1:M)} = w_{1:K}'^{(1:M)}$ with probability

$$A = \min\left\{1, \frac{p(\varepsilon'_L, \varepsilon'_R)p(\theta')q(\varepsilon_L, \varepsilon_R|\varepsilon'_L, \varepsilon'_R)q(\theta|\theta')}{p(\varepsilon_L, \varepsilon_R)p(\theta)q(\varepsilon'_L, \varepsilon'_R|\varepsilon_L, \varepsilon_R)q(\theta'|\theta)} \prod_{k=1}^{K} \frac{\sum_{i=1}^{M} w_k'^{(i)}}{\sum_{i=1}^{M} w_k^{(i)}}\right\}.$$

Otherwise, keep $(\varepsilon_L, \varepsilon_R), \theta$ and set $\bar{w}_{1:K}^{(1:M)} = w_{1:K}^{(1:M)}$.

for $k = 1, \ldots, K$ **do**

Sample $j \in \{1, \ldots, M\}$ w.p. $\propto \bar{w}_k^{(j)}$ and set $(\alpha_k, \beta_k) = (\alpha_k^{(j)}, \beta_k^{(j)})$.

Store $(\varepsilon_L^{(t)}, \varepsilon_R^{(t)}) = (\varepsilon_L, \varepsilon_R)$, $\theta^{(t)} = \theta$.

(α_k, β_k), we used

$$Q_k(\alpha_k^{(i)}, \beta_k^{(i)}) = \text{Beta}(\alpha_k^{(i)}|A_k + 1, n_{0,k} - A_k + 1) \cdot \text{Beta}(\beta_k^{(i)}|B_k + 1, n_{1,k} - B_k + 1).$$

To estimate the ε_{MP} parameter of ε_{MP}-MP and the ε_R parameter of ε_R-BMP-R, we ran Algorithm 2 for $T = 10000$ iterations, and the last 5000 samples are considered after a burn-in period of 5000 iterations.

To jointly estimate the $\varepsilon_L, \varepsilon_R$ parameters of $(\varepsilon_L, \varepsilon_R)$-BMP, we ran Algorithm 2 for $T = 50000$ iterations, and the last 25000 samples are considered after a burn-in period of 25000 iterations.

A.8 Extensions to the MBP estimation method

A.8.1 Modeling attack strength

The model in (5) neglects any conditional dependency among the error probabilities $(\alpha_k, \beta k)$ across k. Also, it assumes that the MIA performance is uniformly distributed, somewhat assigning weak and strong tests equal probability. One can allow for dependence among the error probabilities is to assume a variable $s \in [0, 1]$, with a prior $p(s)$, in the spirit of Yıldırım et al. (2026), and let

$$(\alpha_k, \beta_k)|\varepsilon, s \overset{\text{i.i.d.}}{\sim} \text{Unif}(\mathcal{R}_{\eta_{\nu_k}}(\varepsilon_L, \varepsilon_R) - \mathcal{R}_{(\eta_{\nu_k})^s}(s\varepsilon_L, s\varepsilon_R))$$

Note that in this case (α_k, β_k), $k = 1, \ldots, K$, are dependent (after integrating out s) conditional on $\varepsilon_L, \varepsilon_R$. Moreover, the parameter s can be interpreted as a *attack strength* parameter: The case

$s = 0$ reduces to (5) where the error probabilities are expected to be anywhere in $\mathcal{R}_{\eta_{\nu_j}(\varepsilon_L,\varepsilon_R)}$, thus allowing for weakest tests. As s gets larger, the error probabilities are distributed in a narrower region, farther from the $\alpha + \beta = 1$ line where the worst tests lie.

A.8.2 CORRELATED MIA DECISIONS

In the experiments, for each k, the error counts A_k, B_k for a node are considered Binomial and independent. However, the leave-one-out protocol in Algorithm 1 in Section 4.1 reuses the same pool of shadow models across many challenges, and hence the resulting error counts can be correlated. In that case, g_θ can instead be chosen as a dependence-aware bivariate normal approximation, generalizing the model of Yıldırım et al. (2026); see (9). Consequently, the independent-binomial model for (A_k, B_k) can understate uncertainty. Following the dependence-aware model of Yıldırım et al. (2026), we can use a bivariate normal approximation for the conditional distribution g_θ in (6). For each node k, let $\theta = (\omega_0, \omega_1, \rho)$, where ω_0 and ω_1 capture within-hypothesis decision dependence and ρ captures cross-dependence between false-positive and false-negative counts. Then,

$$g_\theta(a_k, b_k|\alpha_k, \beta_k) = \mathcal{N}\left(\begin{bmatrix} a_k \\ b_k \end{bmatrix}; \begin{bmatrix} n_{0,k}\alpha_k \\ n_{1,k}\beta_k \end{bmatrix}, \Sigma_k\right), \tag{9}$$

where the covariance is given by

$$\Sigma_k = \begin{bmatrix} \alpha_k(1-\alpha_k)\{n_{0,k} + n_{0,k}(n_{0,k}-1)\omega_0\} & n_{0,k}n_{1,k}\rho\sqrt{\alpha_k(1-\alpha_k)\beta_k(1-\beta_k)} \\ n_{0,k}n_{1,k}\rho\sqrt{\alpha_k(1-\alpha_k)\beta_k(1-\beta_k)} & \beta_k(1-\beta_k)\{n_{1,k} + n_{1,k}(n_{1,k}-1)\omega_1\} \end{bmatrix}$$

An independent bivariate model is recovered as the special case $\omega_0 = \omega_1 = \rho = 0$. In the Bayesian estimator, the dependence parameters can be fixed from diagnostics or estimated jointly with $(\varepsilon_L, \varepsilon_R)$.

To reduce the dependency between the decisions, one can also decouple the models used to form challenges from those used to fit the member/non-member hypotheses; this alternative is shown in Algorithm 3.

Algorithm 3 Node-MIA: Alternative 2

for $j = 1, \ldots, J^{(1)}$ **do**
  $S_j^{(1)} \sim \mathcal{P}(\mathcal{G})$, $\Phi_j^{(1)} = \Phi(\mathcal{G}_{S_j^{(1)}}, \mathbf{X}_{S_j}, \mathbf{Y}_{S_j})$. ▷ Sample $J^{(1)}$ datasets and train a GNN on them.
for $j = 1, \ldots, J^{(2)}$ **do**
  $S_j^{(2)} \sim \mathcal{P}(\mathcal{G})$, $\Phi_j^{(2)} = \Phi(\mathcal{G}_{S_j^{(2)}}, \mathbf{X}_{S_j^{(2)}}, \mathbf{Y}_{S_j^{(2)}})$. ▷ Sample $J^{(2)}$ datasets and train a GNN on them (used to challenge MIA)
for $k = 1, \ldots, K$ **do**
  Set $\nu = \nu_k$
  $\hat{\gamma}_\nu(1) = \#\{j \in \{1, \ldots, J^{(1)}\} : \nu \in S_j^{(1)}\}/J^{(1)}$. ▷ Estimate the prior membership probability
  $\mathcal{H}_0 = \{\Phi_j^{(1)} : j \in \{1, \ldots, J^{(1)}\}, \nu \notin S_j^{(1)}\}$, $\mathcal{H}_1 = \{\Phi_j^{(1)} : j \in \{1, \ldots, J^{(1)}\}, \nu \in S_j^{(1)}\}$, ▷ Construct the sets for H_0, H_1
  Set $A_k = B_k = 0$, $n_{0,k} = n_{1,k} = 0$
  for $j \in \{1, \ldots, J^{(2)}\}$ **do**
    $m = \mathbb{I}(\nu \in S_j^{(2)})$. ▷ True membership
    $n_{m,k} \leftarrow n_{m,k} + 1$ ▷ True count
    $\hat{m} \leftarrow$ DECISION$(\mathcal{I}_\nu, y_\nu, \mathcal{H}_0, \mathcal{H}_1, \hat{\gamma}_\nu(1))$ ▷ Decide on the membership
    $A_k \leftarrow A_k + \mathbb{I}(\hat{m} \neq m = 0)$ ▷ type-I error
    $B_k \leftarrow B_k + \mathbb{I}(\hat{m} \neq m = 1)$ ▷ type-II error

A.8.3 AVAILABILITY OF THE PRIOR MEMBERSHIP PROBABILITIES

The joint probability model involves the true membership priors γ_k through $h_{\eta_k}(\alpha_k, \beta_k|\varepsilon_L, \varepsilon_R)$, where $\eta_k = \gamma_k/(1-\gamma_k)$. However, γ_k's are seldom available under a non-trivial sampling mechanism such as snowball sampling.

In the experiments, we used $\widehat{\gamma}_k = n_{1,k}/J$ instead of the ground truths. When the number of shadow models J is large, the estimator has variance $\gamma_k(1-\gamma_k)/J \leq 0.25/J$, which is arguably small enough not to harm the accuracy significantly. However, the discrepancy between the empirical frequency and the ground truth may be important, especially for small J. The discrepancy can be addressed by treating γ_k as unknown with an uninformative prior, and changing the joint model with respect to (7) as

$$p(\varepsilon_L, \varepsilon_R, \theta, \gamma_{1:K}, \alpha_{1:K}, \beta_{1:K}, a_{1:K}, b_{1:K}) = p(\varepsilon_L, \varepsilon_R)p(\theta) \prod_{k=1}^{K} p(\gamma_k)\text{Binom}(n_{1,k}|J, \gamma_k)$$
$$\cdot \prod_{k=1}^{K} h_{\gamma_k/(1-\gamma_k)}(\alpha_k, \beta_k|\varepsilon_L, \varepsilon_R) g_\theta(a_k, b_k|\alpha_k, \beta_k).$$

Sampling from the posterior distribution according to this modified joint model can also be done by slightly modifying MHAAR in Algorithm 2. The required modification involves treating the node-specific γ_ks similarly to the pairs (α_k, β_k)s.

A.8.4 Defense mechanisms with analytical guarantee

In our experiments, we estimate BMP parameters numerically using posterior sampling based on the hierarchical model developed in Section 4.2. An interesting line of future work is developing mechanisms that provide BMP with tight analytical guarantees, although that typically results in very loose bounds. A very recently emerging work (Bai et al., 2026) on the composition of Pufferfish privacy algorithms can be exploited to build complex BMP algorithms from existing mechanisms.

A.9 Datasets and Models

A.9.1 Datasets

For the first three RQs we use citation network datasets from the Planetoid benchmark: Cora and CiteSeer. Cora consists of $2,708$ scientific publications (nodes) classified into 7 research topics (classes). The citation network contains $5,429$ edges. Each node is represented by a 1,433-dimensional sparse bag-of-words feature vector indicating the presence of specific terms in the document. CiteSeer contains $3,327$ scientific publications (nodes) categorized into 6 research areas (classes), with $4,732$ citation links (edges). Node features are $3,703$-dimensional sparse bag-of-words vectors derived from document text. For RQ4 we use Reddit a node-classification benchmark containing 232,965 posts with 602-dimensional features and 41 community labels. Each shadow model is trained on 38,358 nodes, approximately 25% of the official training split, selected using snowball sampling.

A.9.2 Gcn, Gat and Gap models

For the Graph Convolutional Network (GCN) (Kipf & Welling, 2017), the aggregation corresponds to normalized neighborhood feature averaging. Let $\tilde{\mathcal{A}} = \mathcal{A} + \mathbf{I}$ be the adjacency matrix with added self-loops and $\tilde{\mathbf{D}}$ the corresponding degree matrix. The layer-wise update can be written as

$$\mathbf{z}_i^{(\ell)} = \sum_{j \in \mathcal{N}(i)\cup\{i\}} \frac{1}{\sqrt{\tilde{d}_i \tilde{d}_j}} \mathbf{x}_j^{(\ell-1)}, \;\; \mathbf{x}_i^{(\ell)} = \sigma\Big(\mathbf{W}^{(\ell)} \mathbf{z}_i^{(\ell)}\Big),$$

where $\tilde{d}_i$ is the degree of node i in $\tilde{\mathcal{A}}$, $\mathbf{W}^{(\ell)}$ is a learnable weight matrix, and σ is a non-linear activation function. For the Graph Attention Network (GAT) (Veličković et al., 2017), aggregation is performed using learned attention weights that assign different importance to neighboring nodes. For each attention head h, let $\alpha_{ij}^{(\ell,h)}$ denote the normalized attention coefficient from node j to node i. The head-specific representation is $\mathbf{z}_i^{(\ell,h)} = \sum_{j \in \mathcal{N}(i)\cup\{i\}} \alpha_{ij}^{(\ell,h)} \mathbf{W}^{(\ell,h)} \mathbf{x}_j^{(\ell-1)}$. The outputs of the H attention heads are then concatenated to obtain $\mathbf{x}_i^{(\ell)} = \|_{h=1}^{H} \sigma\Big(\mathbf{z}_i^{(\ell,h)}\Big)$ where $\mathbf{W}^{(\ell,h)}$ is a learnable projection matrix and $\|$ denotes concatenation.

Our GRAPHSAGE (Hamilton et al., 2017) implementation has two mean-aggregation layers, a hidden dimension of 32, ReLU activation, and dropout of 0.5. We train for 20 epochs using Adam with learning rate 0.01, weight decay 5×10^{-4}, and cross-entropy loss. Training uses batches of 1,024 seed nodes with neighbor-sampling fanouts of $(15, 10)$. Inference uses all neighbors through layer-wise batches of 4,096 nodes.

GAP (Sajadmanesh et al., 2023) employs a privacy-preserving GNN training procedure that protects node information by privatizing the neighborhood aggregation stage. It first pre-trains an encoder model to obtain node representations, then normalizes these encoded features before computing multi-hop neighborhood aggregations. Instead of using the raw aggregated representations directly, GAP adds calibrated Gaussian noise to the aggregated features at each hop, producing private neighborhood summaries. These private aggregations are then used to train the final classification model, which combines information from different hop distances through base models, a combination module, and a prediction head.

A.9.3 HYPERPARAMETER SETTINGS

For the GCN experiments, we use a 2-layer GCN with hidden dimension 16, trained for 100 epochs. For the GAT experiments, we use a 2-layer GAT trained for 100 epochs, with a hidden dimension of 4 per attention head in the first layer. We used eight attention heads in the first layer and one in the last layer. In the graph sampling pipeline, we set the maximum number of sampled neighbors to 5 (for snowball expansion). For the noisy-output setting, we add Gaussian noise to the parameters of the last layer with standard deviation 0.25. We use the original GAP implementation for our inductive setting with a privacy budget set to 2. The encoder and classification modules are trained for 100 epochs.

A.10 DETAILED EXPERIMENTAL RESULTS

A.10.1 ADDITIONAL RESULTS FOR ESTIMATING MP AND BMP-R PARAMETERS

Tables 2, 3, 4, 5 in Appendix A.10 show the 5th, 50th (median), and 95th percentile (p5 / p50 / p95) estimates of the privacy parameters for MP and BMP-R with CORA with 25% and 50% sampling rates and CITESEER with 25% and 50% sampling rates, respectively.

Table 2: Hierarchical results table (p5 / p50 / p95 per cell) when CORA is employed with 25% nodes in train set.

Sampling	Decision	Output	GCN				GAT			
			full edge		no edges		full edge		no edges	
			BMP-R	MP	BMP-R	MP	BMP-R	MP	BMP-R	MP
random	weak	clean	2.2/2.4/2.5	3.5/3.6/3.6	6.2/7.6/10	7/8.3/11	2.3/2.3/2.4	3.7/3.7/3.7	6.1/7.5/10	7.3/8.6/11
		noisy	1.5/1.6/1.6	2.7/2.8/2.8	3.8/4.3/5.2	4.8/5.1/5.5	1.3/1.3/1.4	2.8/2.8/2.9	3.8/4.1/4.7	4.7/5/5.3
	strong	clean	4.2/5.6/8.5	4.2/4.5/4.8	6.6/8.1/11	7.5/8.8/11	3/3.4/4.2	3.7/3.8/3.8	6.9/8.3/11	7.9/9.2/12
		noisy	2.7/3.4/6.7	3/3.1/3.2	4.8/5.6/7.7	5.3/5.7/6.3	2/2.2/2.4	2.9/2.9/3	4.7/5.3/6.8	5.3/5.7/6.1
snowball	weak	clean	4.4/4.8/5.7	9.7/11/13	6.1/7.4/9.8	9.8/11/13	4.8/5.4/6.7	9.4/11/13	5.1/5.9/8.2	9.5/11/13
		noisy	4.1/4.4/5	9.2/11/13	5.3/6.2/8.3	9.2/11/13	4.5/5.1/6.3	9.3/11/13	5.1/5.9/7.6	9.2/11/13
	strong	clean	5.1/6/8.5	9.9/11/13	5.5/6.6/9.2	10/11/13	5.2/6.4/8.9	9.6/11/13	5.6/6.8/9.4	9.6/11/13
		noisy	4.2/4.7/5.6	9.4/11/13	4.7/5.4/7.4	9.5/11/13	4.6/5.3/7.7	9.6/11/13	4.9/5.8/8.1	9.5/11/13

A.10.2 ILLUSTRATIVE EXAMPLE COMPARING BMP AND MP

Figure 6 shows that several nodes requiring a large MP bound have low membership-prior odds. For this Cora experiment, the largest finite MP bound exceeds the largest finite BMP-R bound.

A.10.3 RESULTS ON AUDITING DP-MODELS.

In addition, we apply our framework to audit the differential private GNN model GAP (Sajadmanesh et al., 2023) trained for *node-level differential privacy* (different from membership differential privacy) with a privacy budget $\varepsilon = 2$.

Table 3: Hierarchical results table (p5 / p50 / p95 per cell) when CORA is employed with 50% nodes sampled to build train graphs.

Sampling	Decision	Output	GCN				GAT			
			full edge		no edges		full edge		no edges	
			BMP-R	MP	BMP-R	MP	BMP-R	MP	BMP-R	MP
random	weak	clean	4/4.3/4.7	4.3/4.6/4.9	6.6/7.9/10	6.8/8.1/10	3/3.1/3.2	3.4/3.4/3.4	6.3/7.7/10	6.4/7.6/10
		noisy	2.6/2.6/2.6	3/3/3.1	4.3/4.6/5.1	4.4/4.6/4.9	2.7/2.8/2.8	3/3/3	4.5/4.9/5.4	4.5/4.7/5
	strong	clean	3.8/4.1/4.4	4/4.1/4.3	6.6/7.9/10	6.7/8/11	3.2/3.3/3.4	3.4/3.5/3.5	6.3/7.7/10	6.3/7.4/10
		noisy	2.5/2.6/2.6	3.1/3.1/3.1	4.4/4.7/5.2	4.4/4.6/4.9	2.7/2.7/2.8	3/3/3.1	4.6/4.9/5.5	4.5/4.8/5.1
snowball	weak	clean	7.1/8.5/11	10/11/13	7.1/8.4/11	10/11/13	7.2/8.5/11	10/11/13	5.9/6.6/8.5	9.8/11/13
		noisy	5.7/6.3/8.3	9.5/11/13	6.4/7.5/9.6	9.5/11/13	6.1/7/8.9	9.7/11/13	5.7/6.3/7.6	9.6/11/13
	strong	clean	6.8/8/10	10/12/14	6.7/7.9/10	10/11/13	6.7/8/10	10/12/14	5.8/6.4/8.3	10/11/13
		noisy	5.6/6.3/8.1	9.8/11/13	5.5/6.1/7.3	9.9/11/13	5.8/6.6/8.2	10/11/14	5.5/6/7.3	9.8/11/13

Table 4: Hierarchical results table (p5 / p50 / p95 per cell) when CITESEER is employed with 25% nodes sampled to build train graphs.

Sampling	Decision	Output	GCN				GAT			
			full edge		no edges		full edge		no edges	
			BMP-R	MP	BMP-R	MP	BMP-R	MP	BMP-R	MP
random	weak	clean	6.1/7.5/9.9	7.6/8.9/11	7.3/8.7/11	8.5/9.9/12	5.9/7.3/9.9	7/8.2/10	7.5/8.9/11	9/10/13
		noisy	3.1/3.3/3.5	4.1/4.2/4.3	4.8/5.6/7.7	5.7/6.1/7	2.9/3/3.2	4.1/4.2/4.2	4.3/4.8/6.2	5.4/5.8/6.6
	strong	clean	6.8/8.3/11	7.9/9.2/12	8/9.4/12	9.1/10/12	6.6/8.1/11	7.5/8.9/11	8.2/9.6/12	9.3/11/13
		noisy	4.4/5.4/8	4.8/5/5.3	6/7.4/10	6.6/7.4/9.3	4/4.8/7.1	4.4/4.6/4.9	5.5/6.5/9	6/6.4/7.6
snowball	weak	clean	5.1/6.5/9.1	11/13/14	5.5/6.9/9.6	11/13/15	3.8/4.3/5.7	11/12/14	3.5/3.8/4.4	11/12/14
		noisy	3.5/4/4.8	11/12/14	3.7/4/4.8	11/12/14	2.9/3.1/3.4	11/12/14	2.9/3/3.2	11/12/14
	strong	clean	6.3/7.7/10	11/13/15	6.3/7.7/10	12/13/14	5.4/6.9/9.7	11/13/15	4.9/6.3/9.1	11/13/14
		noisy	4.8/6.3/9.1	11/13/15	5/6.5/9.2	11/13/14	4.6/6.1/9.1	11/12/14	3.8/5.3/8	11/12/14

Table 6 compares the BMP estimates for GAP, trained in an inductive setting with a privacy budget of $\varepsilon = 2$, against non-private GCN and GAT models on Cora with a 50% node split. We use this setting because Cora is denser than CITESEER, making neighborhood aggregation meaningful and allowing structural effects in graph learning to be more clearly reflected in the audit. GAP generally yields much lower BMP-R values than GCN and GAT, showing that the DP mechanism reduces membership leakage. In addition, once again we see that both MP and BMP estimates are sensitive to the graph-sampling strategy in all algorithms: For example, under GAP, the BMP-R

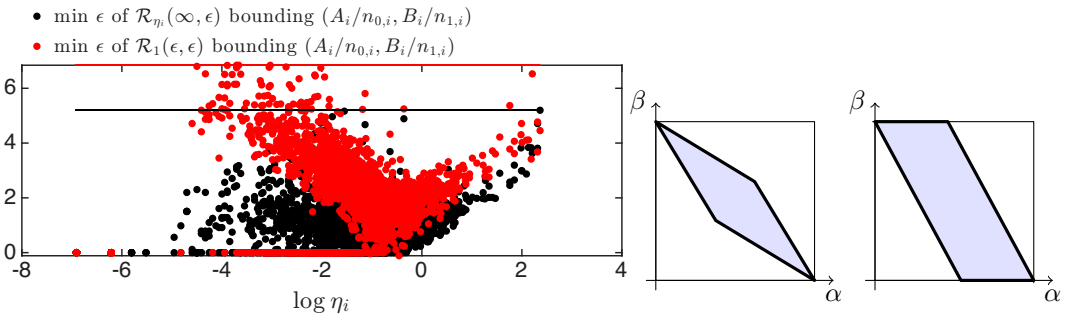


Figure 6: **Left:** Minimum per-node privacy bounds for empirical error rates in the Cora experiment with a 25% sampling rate, GCN, clean outputs, full edges, and a strong attack. The horizontal axis is the log membership-prior odds. Black dots show the minimum ε_R for $\mathcal{R}_{\eta_i}(\infty, \varepsilon_R)$; red dots show the minimum ε_{MP} for $\mathcal{R}_1(\varepsilon_{\text{MP}}, \varepsilon_{\text{MP}})$. Horizontal lines mark the largest finite bounds. **Right:** Example MP and BMP-R error regions.

Table 5: Hierarchical results table (p5 / p50 / p95 per cell) when CITESEER is employed with 50% nodes sampled to build train graphs.

Sampling	Decision	Output	GCN				GAT			
			full edge		no edges		full edge		no edges	
			BMP-R	MP	BMP-R	MP	BMP-R	MP	BMP-R	MP
random	weak	clean	7.3/8.6/11	8.3/9.6/12	8.4/9.9/12	8.9/10/12	6.6/8/10	7.1/8.5/11	7.7/9/11	8/9.4/11
		noisy	4.3/4.6/5	4.5/4.7/5	5.7/6.3/7.8	5.7/6.1/6.9	3.5/3.6/3.8	4.2/4.3/4.3	4.6/4.9/5.3	4.5/4.7/5
	strong	clean	7.2/8.4/11	8.2/9.7/12	8.5/9.9/12	8.8/10/12	6.6/8/10	7.1/8.4/11	7.6/9/11	8.1/9.3/12
		noisy	4.3/4.6/5	4.6/4.8/5.1	5.6/6.3/8	5.6/6.1/7	3.6/3.8/3.9	4.1/4.1/4.2	4.6/4.9/5.3	4.6/4.7/5
snowball	weak	clean	8.7/10/12	12/13/15	7.3/8.8/11	11/13/15	7.1/8.5/11	12/13/15	5.5/6.2/8.4	11/13/14
		noisy	5.6/6.3/8.3	11/12/14	5/5.4/6.6	11/12/14	4.6/4.9/5.4	11/12/14	4.3/4.5/4.9	11/12/14
	strong	clean	9/10/13	12/13/15	7.4/8.8/11	11/13/15	7.3/8.5/11	12/13/15	5.6/6.5/8.8	11/13/15
		noisy	5.7/6.6/8.8	11/13/14	5.1/5.6/7	11/13/14	4.7/5.1/6.2	11/13/15	4.3/4.5/4.9	11/12/14

Table 6: p5/p50/p95 for CORA. Sampling rate for trainng: 50%. Attack: *clean*, *strong*, *full-edges*

Privacy	GAT-random	GCN-random	GAP-random	GAT-snowball	GCN-snowball	GAP-snowball
BMP-R	3.2/3.3/3.4	3.8/4.1/4.4	1.4/1.4/1.4	6.7/8/10	6.8/8/10	3.4/3.4/3.5
MP	3.4/3.5/3.5	4/4.1/4.3	2.3/2.3/2.4	10/12/14	10/12/14	9.4/11/13

median increases from 1.4 with random sampling to 3.4 with snowball sampling, even though the training mechanism and privacy budget are unchanged. This shows that graph-dependent sampling introduces structure-level membership distinguishability that is captured by BMP but not by the DP budget of GAP alone.

The results also provide a practical demonstration that the DP budget cannot be directly lower-bounded from MIA success in GNNs. MP is the closest analogue to the standard DP-auditing setting because it does not condition on heterogeneous prior membership probabilities. Yet, for GAP trained with a nominal privacy budget of $\varepsilon = 2$, MP produces substantially larger estimates, especially under snowball sampling. Finally, GAP produces considerably tighter posterior intervals than GCN and GAT. Since BMP is estimated from attacker error counts, these narrower intervals indicate more stable attack behavior, plausibly because the noise injected by GAP suppresses model-specific variation and reduces member–non-member distinguishability.

A.10.4 EFFECT OF NUMBER OF SHADOW MODELS ON MEMBERSHIP PRIORS AND ATTACK RATES

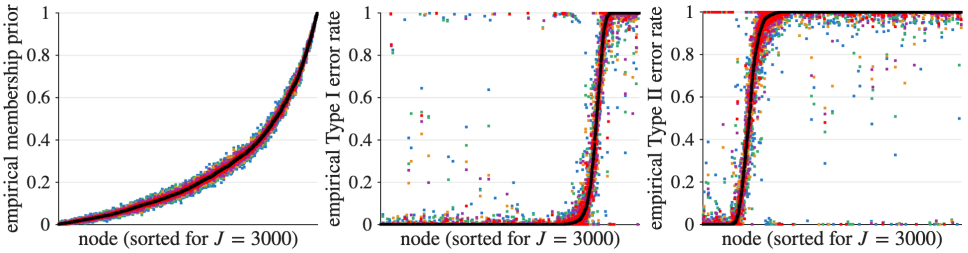


Figure 7: Empirical membership priors and attack error rates as the number J of shadow models increases. Blue, green, orange, purple, red, and black markers denote $J = 400, 600, 800, 1000, 2000, 3000$, respectively. The empirical rates converge as J increases.

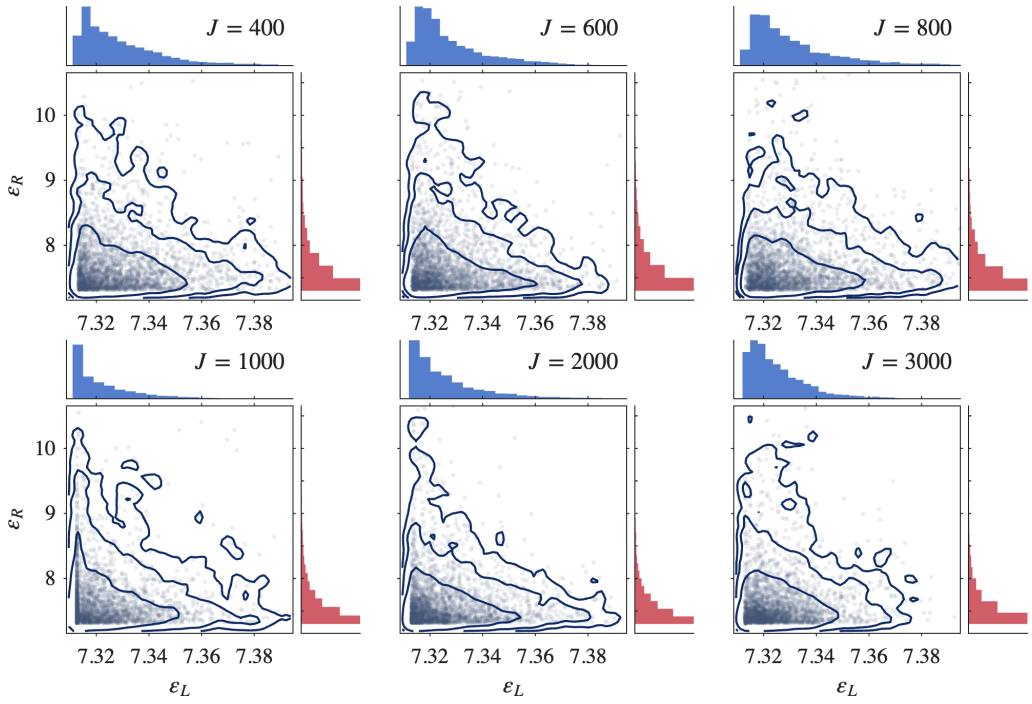


Figure 8: Posterior distribution for varying number of shadow models $J = 400, 600, 800, 1000, 2000, 3000$. The empirical rates converge as J increases.