Imprecise Probabilities for Privacy–Accuracy Trade-Offs in Bayesian Networks

Niccolò Rocchi[1,2] **Fabio Stella**[1] **Cassio de Campos**[3]

[1]University of Milano-Bicocca, Milan, Italy
[2]Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[3]Eindhoven University of Technology, Eindhoven, The Netherlands

Abstract

Bayesian networks are well-known probabilistic graphical models that enable explainable reasoning under uncertainty. In many domains, data are scarce and fragmented across institutions, motivating collaboration and the sharing of learned models rather than individual-level records to increase the available evidence and reduce bias. Releasing a model, however, can still reveal sensitive information: adversaries may mount membership inference attacks to determine whether a specific individual contributed to the training. A common mitigation strategy is to inject noise into the learned model before its release. This perturbation may compromise the quality and interpretability of subsequent inferences. We investigate how a Bayesian network can be effectively masked rather than perturbed to protect it against attacks by leveraging credal networks, the imprecise version of Bayesian networks, allowing us to employ sets of parameters instead of exact point estimates. We formalize and compare various masking and attack strategies, and investigate how privacy leakage depends on them. Finally, we analyze how privacy and utility can be traded off in credal naive Bayes models, comparing them with the common noise-injection baseline. Results indicate that credal masking provides principled protection, substantially reducing attack success while yielding fine-tuned privacy-accuracy trade-offs.

1 INTRODUCTION

Bayesian networks (BN) [Koller and Friedman, 2010] are widely used models for transparently representing multivariate probability distributions in complex systems and reasoning under uncertainty. Formally, a BN is defined by (i) a directed acyclic graph whose *nodes* are random variables and whose *edges* encode conditional independences/dependencies between them, and (ii) a set of parameters that compactly specify a joint distribution over all variables. When edges are interpreted as cause-and-effect relationships between variables, a BN provides rich insights into the data-generating process, supporting causal and counterfactual reasoning [Pearl, 1995]. These reasons have encouraged a broad adoption of these models, especially in high-stakes domains where informed decision-making is essential, including healthcare [Grube et al., 2022, Roy et al., 2023, Suter et al., 2022].

Learning a BN typically involves recovering its structure and estimating its parameters [Scanagatta et al., 2019]. In practice, this can be done using data alone, expert knowledge, or a combination of both [Constantinou et al., 2023]. However, in sensitive domains, data are often scarce (e.g., rare diseases or time-to-event data) [Rocchi et al., 2026], which recently motivated multi-institutional collaborations to increase sample sizes and reduce selection bias [Zanga et al., 2023]. For privacy and governance reasons, however, institutions may not be able to share individual-level data, even after de-identification, and may instead share learned models [Torrijos et al., 2025]. Nevertheless, releasing a BN can still leak information about the individuals used for training: *membership inference attacks* (MIA) from adversaries exploit the released model to determine whether a candidate data record was in the training set [Carlini et al., 2022]. This raises the question of how to release a BN model while minimizing privacy losses.

A standard defense to MIA is to release a *noisy* BN, i.e., a BN obtained by perturbing the learned parameters with random noise, often motivated by the *differential privacy* framework [Dwork et al., 2017]. While effective in principle, perturbations may degrade the model's utility and interpretability and introduce probabilistic inconsistencies when not carefully handled [Binkyte et al., 2024]. Rocchi et al. [2025] recently proposed *credal networks* (CN) [Cozman, 2000] as a masking strategy for BNs. In a CN, each local

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:5716–5730.

conditional distribution is replaced by a *credal set* (a set of plausible probability distributions), thereby representing epistemic uncertainty and yielding *guaranteed* inferential bounds via diverse algorithmic approaches [Antonucci et al., 2015, de Campos and Cozman, 2007]. The core insight of this procedure is that expanding a BN into sets of plausible parameter values can reduce the information available to MIA while preserving the model's semantics, keeping it close to that of the original BN. Moreover, empirical evidence has shown that the effectiveness of MIA depends on the degree of uncertainty in the CN.

However, existing evidence is limited: prior work primarily considers a *single* attack and a *single* defense mechanism, i.e., one fixed choice of how the adversary scores membership and one fixed choice of how the BN is masked (equivalently, how a CN is learned from data) [Rocchi et al., 2025]. As a consequence, it remains unclear whether the observed privacy qualities of a CN persist when the attacker or the masking mechanisms change. Moreover, there is a lack of quantitative evidence demonstrating that a CN offers a fairer privacy-accuracy trade-off than a noisy BN. Qualitative motivation is instead clear: while the inferential bounds generated by the CN may sometimes be broad, leading to "I do not know" conclusions, they still ensure correct responses to queries, unlike noisy models. We study multiple attack and defense mechanisms, and we test whether privacy can still be fine-tuned consistently. Then, we quantify the privacy-accuracy trade-off in controlled *credal naive Bayes* (CNB) [Zaffalon, 2002] testbeds, using both synthetic and real-world data. For each target privacy level, we calibrate a CN and a noisy BN to achieve *matched privacy levels* (i.e., the same privacy score under our MIA evaluation protocol) and then compare their classification accuracy.

2 PROBABILISTIC GRAPHICAL MODELS

Probabilistic graphical models [Koller and Friedman, 2010] offer a principled approach to encoding and managing high-dimensional probability distributions. This article focuses on *Bayesian networks* (BN) and their extension as *credal networks* (CN).

BNs and CNs rely on a *graph* $\mathcal{G}$, i.e., a mathematical object consisting of a set of nodes $\mathbf{X}$ and edges connecting pairs of nodes. If $X \to Y$ in $\mathcal{G}$ then X is said to be a *parent* of Y, and Π_Y denotes the set of parents of Y. A *directed acyclic graph* (DAG) is a graph in which all edges are directed ($X \to Y$) and there is no sequence of nodes $X \to \cdots \to Y$ such that $X = Y$. We consider DAGs whose nodes are associated with random variables; namely, $X \in \mathbf{X}$ denotes both a node in $\mathcal{G}$ and a random variable with a given distribution. In this paper, we focus on categorical variables. Hence, we denote by $\mathrm{Val}(X)$ the set of values that X may attain; similarly, we define $\mathrm{Val}(\mathbf{X}) = \times_{X \in \mathbf{X}} \mathrm{Val}(X)$.

2.1 BAYESIAN NETWORKS

Definition 1 (BN)**.** *A BN is a pair $(\mathcal{G}, \boldsymbol{\theta})$ where $\mathcal{G}$ is a DAG over random variables $\mathbf{X}$ and $\boldsymbol{\theta}$ is a set of parameters defining a joint distribution over $\mathbf{X}$ that factorizes according to $\mathcal{G}$:*

$$P(\mathbf{X}; \boldsymbol{\theta}) = \prod_{X \in \mathbf{X}} P(X \mid \Pi_X; \boldsymbol{\theta}_{X|\Pi_X}) \,. \tag{1}$$

In Equation (1), $\boldsymbol{\theta}_{X|\Pi_X}$ defines the *local* conditional distribution of X and is part of defining $\boldsymbol{\theta}$. In other words, $X \mid \Pi_X \sim \mathrm{Multinomial}(\boldsymbol{\theta}_{X|\Pi_X})$ and the joint distribution is the product of these multinomials. In particular, $\boldsymbol{\theta}_{X|\Pi_X} = \{\boldsymbol{\theta}_{X|\pi_X} : \pi_X \in \mathrm{Val}(\Pi_X)\}$, so we have a conditional distribution for any value of the parents of X. This distribution is specified by a set of point parameters, namely $\boldsymbol{\theta}_{X|\pi_X} = \{\theta_{x|\pi_X} : x \in \mathrm{Val}(X)\}$, where $\theta_{x|\pi_X}$ defines the conditional probability $P(x \mid \pi_X)$. Hereafter, let Θ denote the entire family of parameter sets $\boldsymbol{\theta}$ whose distribution factorizes as in Equation (1).

The *complexity* of a BN, denoted by $C(\mathcal{G})$, is the total number of free parameters of its nodes, given by:

$$C(\mathcal{G}) = \sum_{X \in \mathbf{X}} |\mathrm{Val}(\Pi_X)| \cdot (|\mathrm{Val}(X)| - 1) \,. \tag{2}$$

2.2 CREDAL NETWORKS

CNs [Cozman, 2000] are models based on Walley's imprecise probability theory [Walley, 2000] and represent uncertainty in the BN's parameters.

Definition 2 (Credal set)**.** *Let X be a random variable. A credal set $\mathcal{K}(X)$ is a closed and convex subset of the space of probability distributions over* $\mathrm{Val}(X)$*. A joint credal set $\mathcal{K}(\mathbf{X})$ is defined similarly over* $\mathrm{Val}(\mathbf{X})$.

Let X be parameterized by $\boldsymbol{\theta}_X = \{\theta_{x_1}, \ldots, \theta_{x_k}\}$. Because $\theta_{x_k} = 1 - \sum_{i=1}^{k-1} \theta_{x_i}$, the credal set $\mathcal{K}(X)$ is a subset of the simplex $\Delta^{k-1} \subset [0,1]^k$. With an abuse of notation, we shall write $\mathcal{K}(\boldsymbol{\theta}_X)$ or $\mathcal{K}(X)$ interchangeably, and $\mathcal{K}(\theta_{x_i})$ to mean the projected set into θ_{x_i}. We focus on credal sets defined by intervals, that is, $\mathcal{K}(\theta_{x_i}) = [l_{x_i}, u_{x_i}] \subseteq [0,1]$ for all i. $\mathcal{K}(\boldsymbol{\theta}_X)$ is built of these $\mathcal{K}(\theta_{x_i})$ plus the equality constraint to make them sum 1, assuming coherence.

A *conditional* credal set $\mathcal{K}(X \mid Y)$ is usually defined elementwise, that is, obtained from $\mathcal{K}(X, Y)$ by applying the Bayes rule to each element and possibly taking their convex hull [Cozman and de Campos, 2004]. A collection of conditional credal sets $\mathcal{K}(X \mid Y)$ is said to be *separately specified* when $\mathcal{K}(X \mid Y = y_1)$ and $\mathcal{K}(X \mid Y = y_2)$ are unrelated for any $y_1 \neq y_2$, that is, not constrained by each other. Conditional credal sets behave similarly to non-conditional ones for any value of the conditioning set, and both *H-* and *V-representations* extend naturally [Zaffalon, 2002].

CNs are DAGs associated with conditional credal sets. While a CN can be defined in different ways, we focus on *locally and separately specified* CNs. Other types of CNs are outside the scope of this article and are discussed by Corani and de Campos [2010].

Definition 3 (Locally and separately specified CN)**.** *A locally and separately specified CN consists of a DAG $\mathcal{G}$ where each node X is associated with a local credal set $\mathcal{K}(X \mid \Pi_X)$, and $\mathcal{K}(X \mid \Pi_X)$ is separately specified for any $\pi_X \in \mathrm{Val}(\Pi_X)$.*

Following the above definition, we have that a CN generalizes a BN by replacing each point parameter $\theta_{x|\pi_X}$ with an interval (or, more generally, a credal set). These credal sets can then be combined via the CN *strong extension* [Cozman and de Campos, 2004], which mimics the same probabilistic properties of a BN.

3 PRECISE MEMBERSHIP INFERENCE

Consider a *general population* $\mathcal{P}$ of individuals characterized by a set of variables $\mathbf{X}$, and let $(\mathcal{G}, \boldsymbol{\theta}_0)$ denote the (unknown) BN representing the data-generating process underlying $\mathbf{X}$. Assume that a BN $(\mathcal{G}, \widehat{\boldsymbol{\theta}}_{\mathcal{T}})$ is learned using data on the variables $\mathbf{X}$ from a *target dataset* $\mathcal{T} \subset \mathcal{P}$. In particular, we let $\widehat{\boldsymbol{\theta}}_{\mathcal{T}}$ denote the maximum-likelihood estimate (MLE) of $\boldsymbol{\theta}_0$, obtained from the target data $\mathcal{T}$. Furthermore, assume that the data $\mathcal{T}$ are *private*; hence, only the *data owner* and the *learning algorithm* have direct access to them. Nevertheless, the BN $(\mathcal{G}, \widehat{\boldsymbol{\theta}}_{\mathcal{T}})$ may be learned and then publicly released for research purposes. In the medical domain, for example, $\mathcal{P}$ may represent all patients affected by a particular disease, while $\mathcal{T}$ corresponds to a hospital database containing a subset of those patients. An *attacker* is a malicious actor with access to the released BN $(\mathcal{G}, \widehat{\boldsymbol{\theta}}_{\mathcal{T}})$ who exploits this knowledge to infer private information about individuals in the target data $\mathcal{T}$. In particular, the attacker could leverage both the BN $(\mathcal{G}, \widehat{\boldsymbol{\theta}}_{\mathcal{T}})$ and an auxiliary *reference* data $\mathcal{R} \subset \mathcal{P}$, to conduct a *membership inference attack* (MIA) against the target $\mathcal{T}$. An MIA, or *tracing attack*, is a technique for determining whether a specific individual $\mathbf{x} \in \mathcal{P}$ belongs to the target dataset $\mathcal{T}$ [Carlini et al., 2022]. In the above example, $\mathcal{R}$ could be a publicly available clinical registry or an auxiliary model. Another example may concern a competing insurance company seeking to attract more clients from a rival firm. In this setting, $\mathcal{P}$ denotes a vulnerable population, whereas $\mathcal{T}$ and $\mathcal{R}$ denote the subsets associated with the competing firm and the insurance company, respectively.

In the MIA setting, $(\mathcal{G}, \widehat{\boldsymbol{\theta}}_{\mathcal{T}})$ and $\mathcal{R}$ are exploited by the attacker to build a *decision rule* $\phi(\cdot)$ so that, given $\mathbf{x} \in \mathcal{P}$, it yields either $\phi(\mathbf{x}) = 1$, hence concluding $\mathbf{x} \in \mathcal{T}$, or $\phi(\mathbf{x}) = 0$, hence concluding $\mathbf{x} \notin \mathcal{T}$. Because there is no access to $\mathcal{T}$, a surrogate approach to defining this decision problem is via the parametrization that potentially generated the data. Let $P(\mathbf{X}; \boldsymbol{\theta}_{\mathbf{X}})$ be the distribution from which $\mathbf{x}$ has been generated. We can formulate the following hypothesis test:

$$\begin{cases} H_0 : \ \boldsymbol{\theta}_{\mathbf{X}} = \boldsymbol{\theta}_0 \neq \widehat{\boldsymbol{\theta}}_{\mathcal{T}} & (\phi(\mathbf{x}) = 0)\,, \\ H_1 : \ \boldsymbol{\theta}_{\mathbf{X}} = \widehat{\boldsymbol{\theta}}_{\mathcal{T}} & (\phi(\mathbf{x}) = 1)\,, \end{cases} \tag{3}$$

where $\boldsymbol{\theta}_0$ is associated with the true but unknown BN $(\mathcal{G}, \boldsymbol{\theta}_0)$ generating the population $\mathcal{P}$. This brings us to the simple vs. simple hypotheses of the *log-likelihood ratio* (LLR) test [Lehmann and Romano, 2022]. While we acknowledge alternative approaches for conducting an MIA attack [Zarifzadeh et al., 2024, Carlini et al., 2022], the choice of the LLR function is motivated by: (i) the specific setting where $\mathcal{T}$ has a low sample size; (ii) the natural decomposability of the LLR function with respect to the BN structure.

The LLR statistic compares the log-likelihood of $\widehat{\boldsymbol{\theta}}_{\mathcal{T}}$ versus that of $\boldsymbol{\theta}_0$. When the number of individuals in $\mathcal{R}$ is *sufficiently large*, $\widehat{\boldsymbol{\theta}}_{\mathcal{R}}$ can be considered the MLE estimate of $\boldsymbol{\theta}_0$. For this reason, the employed LLR statistic is commonly:

$$\Lambda : \ \mathrm{Val}(\mathbf{X}) \to \mathbb{R}, \quad \Lambda(\mathbf{x}) = L(\widehat{\boldsymbol{\theta}}_{\mathcal{T}} \mid \mathbf{x}) - L(\widehat{\boldsymbol{\theta}}_{\mathcal{R}} \mid \mathbf{x})\,, \tag{4}$$

where $L(\boldsymbol{\theta} \mid \mathbf{x}) = \log P(\mathbf{x}; \boldsymbol{\theta})$ is the log-likelihood function for $\boldsymbol{\theta} \in \Theta$.

The test rejects H_0 whenever $\Lambda(\mathbf{x})$ is *large enough*, that is, when the former likelihood deviates markedly from the latter, providing sufficient evidence to conclude that $\mathbf{x} \in \mathcal{T}$. Let α be the Type I error rate (i.e., the *significance level*), defined as the probability that the attacker erroneously infers $\mathbf{x} \in \mathcal{T}$ when the true condition is $\mathbf{x} \notin \mathcal{T}$. The attacker specifies the tolerable Type I error level α, and the distribution of $\Lambda(\cdot)$ is estimated under H_0. Next, the threshold $\tau(\alpha)$ is chosen so that $P(\Lambda(\mathbf{x}) > \tau(\alpha) \mid \mathbf{x} \notin \mathcal{T}) = \alpha$. Then, the value of $\Lambda(\mathbf{x})$ is computed and the null hypothesis H_0 is rejected, in favor of H_1 ($\mathbf{x} \in \mathcal{T}$) whenever $\Lambda(\mathbf{x}) > \tau(\alpha)$. Indeed, when $\Lambda(\mathbf{x})$ does not exceed $\tau(\alpha)$, there is insufficient evidence to conclude $\mathbf{x} \in \mathcal{T}$. The attacker also takes into account the *attack power*, defined as $\beta = P(\Lambda(\mathbf{x}) > \tau(\alpha) \mid \mathbf{x} \in \mathcal{T})$. In the simple vs. simple hypotheses problem, the LLR test achieves the highest power among all tests for any value of the Type I error rate α, due to the Neyman-Pearson lemma [Neyman and Pearson, 1933].

Note that $L(\boldsymbol{\theta} \mid \mathbf{x})$ is a sum of log-likelihood ratios for the network's single parameters, which are asymptotically χ^2 with one degree of freedom [Wilks, 1938]. If $|\mathbf{X}|$ is large *enough*, this sum is approximately Normal and can be used to guide the choice of $\tau(\alpha)$ to control Type I error. By leveraging the LLR properties, Kumar Murakonda et al. [2021] claimed a relationship between α and β:

Theorem 3.1. *For any error level α, it holds:*

$$z_\alpha + z_{1-\beta} \approx \sqrt{\frac{C(\mathcal{G})}{|\mathcal{T}|}}, \tag{5}$$

where z_s, $0 < s < 1$, is the quantile at level $1 - s$ of the Standard Normal distribution $\mathcal{N}(0, 1)$.

In the literature, model privacy is commonly characterized by the indistinguishability of an algorithm's output across *neighbor datasets* [Dwork et al., 2017]. Two neighbor datasets may differ by a single addition, removal, or modification of a data item. This justifies the following definition, which is the typically employed notion of algorithmic privacy.

Definition 4 (ϵ-differential privacy)**.** *A randomized algorithm $\mathcal{A}$ satisfies ϵ-differential privacy (ϵ-DP) if, for any neighbor datasets $\mathcal{D}$ and $\mathcal{D}'$, and for any output $\mathcal{O}$, it holds:*

$$\left| \log \frac{P(\mathcal{A}(\mathcal{D}) = \mathcal{O})}{P(\mathcal{A}(\mathcal{D}') = \mathcal{O})} \right| \leqslant \epsilon\,. \tag{6}$$

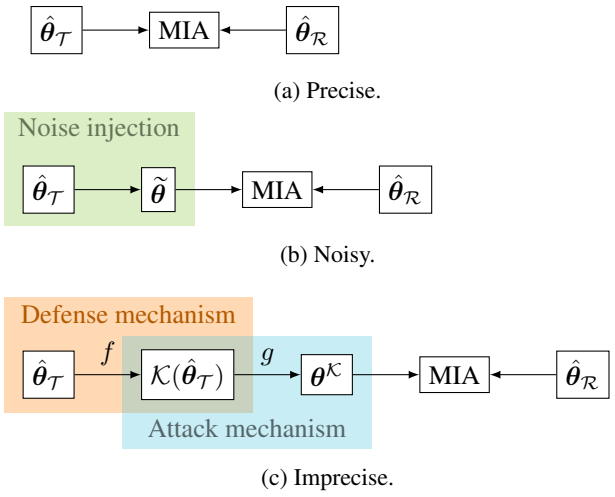
Most techniques that make an algorithm ϵ-DP involve injecting noise into its output. A *noisy* BN $(\mathcal{G}, \widetilde{\boldsymbol{\theta}})$ can be obtained after injecting Laplacian noise into its parameters [Zhang et al., 2017]. The Laplacian perturbation is effective at protecting privacy, but it directly mines the model's informativeness and utility, especially when a BN model is used.

4 IMPRECISE MEMBERSHIP INFERENCE

In the typical MIA-BN setting, the attacker accesses and leverages the unprotected $\widehat{\boldsymbol{\theta}}_{\mathcal{T}}$ (Figure 1a). When noise injection is employed instead, a perturbed version of the BN is released, hindering direct access to the sensitive information held by $\widehat{\boldsymbol{\theta}}_{\mathcal{T}}$ (Figure 1b). This Section explores how BNs can be veiled by imprecise probabilities rather than muddled by noise, thereby enforcing protection against MIA.

4.1 SETTINGS

When a CN is released, the parameters $\widehat{\boldsymbol{\theta}}_{\mathcal{T}}$ are exposed only via their respective credal sets. In doing so, the attacker has access to a non-trivial neighborhood of $\widehat{\boldsymbol{\theta}}_{\mathcal{T}}$ rather than to point parameters. Assuming the attacker uses the same attack framework, they need to choose $\boldsymbol{\theta}^{\mathcal{K}} \in \mathcal{K}(\widehat{\boldsymbol{\theta}}_{\mathcal{T}})$ to execute the attack. We call this setting *imprecise* MIA, as shown in Figure 1c. We refer to the construction of the CN as the *defense mechanism*, implemented on the data side before any model release. By contrast, we refer to the choice of $\boldsymbol{\theta}^{\mathcal{K}} \in \mathcal{K}(\widehat{\boldsymbol{\theta}}_{\mathcal{T}})$ as the *attack mechanism*, which the



(a) Precise.

(b) Noisy.

(c) Imprecise.

Figure 1: The settings considered for MIA; $\widehat{\boldsymbol{\theta}}_{\mathcal{T}}, \widehat{\boldsymbol{\theta}}_{\mathcal{R}}, \widetilde{\boldsymbol{\theta}}$, and $\boldsymbol{\theta}^{\mathcal{K}}$ refer to BNs, while $\mathcal{K}(\widehat{\boldsymbol{\theta}}_{\mathcal{T}})$ refer to a CN.

attacker implements after the CN is released. Being $\wp(\cdot)$ the power set operator, defining a defense mechanism means describing the mapping:

$$f:\ \Theta \to \wp(\Theta), \quad f(\widehat{\boldsymbol{\theta}}_{\mathcal{T}}) = \mathcal{K}(\widehat{\boldsymbol{\theta}}_{\mathcal{T}})\,. \tag{7}$$

Conversely, an attack mechanism defines a mapping:

$$g:\ \wp(\Theta) \to \Theta, \quad g(\mathcal{K}(\widehat{\boldsymbol{\theta}}_{\mathcal{T}})) = \boldsymbol{\theta}^{\mathcal{K}}\,. \tag{8}$$

As a conclusion, the imprecise MIA setting can be shortly described as the combination $(g \circ f)(\widehat{\boldsymbol{\theta}}_{\mathcal{T}})$. Sections 4.2 and 4.3 present different choices for the functions f and g, respectively. For now, we refrain from making any assumptions about the attacker's knowledge of the defense mechanism.

In the imprecise MIA, the LLR statistic depends on the chosen attack mechanism, whereas the test hypotheses remain identical to those in the precise setting (Equation (3)). Hence, Equation (4) becomes:

$$\begin{aligned} \Lambda:\ &\mathrm{Val}(\mathbf{X}) \times \mathcal{K}(\widehat{\boldsymbol{\theta}}_{\mathcal{T}}) \to \mathbb{R}, \\ \Lambda(\mathbf{x}, \boldsymbol{\theta}^{\mathcal{K}}) &= L(\boldsymbol{\theta}^{\mathcal{K}} \mid \mathbf{x}) - L(\widehat{\boldsymbol{\theta}}_{\mathcal{R}} \mid \mathbf{x})\,. \end{aligned} \tag{9}$$

We denote by $\phi(\cdot, \boldsymbol{\theta}^{\mathcal{K}})$ the set of decision rules associated with the imprecise version of the test, with one rule for each $\boldsymbol{\theta}^{\mathcal{K}} \in \mathcal{K}(\widehat{\boldsymbol{\theta}}_{\mathcal{T}})$. Notice that the attacker always knows $\widehat{\boldsymbol{\theta}}_{\mathcal{R}}$. By setting $\boldsymbol{\theta}^{\mathcal{K}} = \widehat{\boldsymbol{\theta}}_{\mathcal{T}}$ in Equation (9), we fall back to the precise MIA setting as in Section 3, though this is not possible for the attacker to do, given their knowledge. Section A provides theoretical insights into the behavior of the statistic in the imprecise case.

The trade-off between the uncertainty inherent in the network and the privacy to be preserved depends on the specific case. Intuitively, larger parameter intervals lead to greater

privacy but also reduce the model's utility. We can mark two extreme situations:

1. $\mathcal{K}(\widehat{\boldsymbol{\theta}}_{\mathcal{T}}) = \{\widehat{\boldsymbol{\theta}}_{\mathcal{T}}\}$. In this case, we release a single BN, and the power of MIA is approximated by Theorem 3.1 [Kumar Murakonda et al., 2021]. Hence, we release an unmasked model, thereby obtaining the minimum privacy.
2. $\mathcal{K}(\widehat{\boldsymbol{\theta}}_{\mathcal{T}}) = \Theta$. In this case, each parameter interval is $[0, 1]$, thus only the DAG is released. Hence, we release the least information while ensuring the highest possible privacy.

4.2 DEFENSE MECHANISMS

We focus on defense mechanisms that yield locally and separately specified CNs. Hence, the function f of the defense mechanism can be defined on each point parameter $\theta_{x_i|\pi_X}$ so that $f(\theta_{x_i|\pi_X}) = [l_{x_i|\pi_X},\ u_{x_i|\pi_X}]$. Surely, setting each $f_{x_i|\pi_X}$ independently of the others could lead to an inconsistent CN, so this should be taken into account.

Imprecise Dirichlet Model Defense The *imprecise Dirichlet model* (IDM) [Walley, 1996] is among the most widely used methods for learning a CN from data.

Assume X has no parents, $|\mathrm{Val}(X)| = k$. In the *precise* setting of the Bayesian posterior estimator, $\boldsymbol{\theta}_X$ has a prior distribution $\mathrm{Dirichlet}(\mathbf{c})$, where $\mathbf{c} = \{c_1, \ldots, c_k\}$ are user-defined so that $c_i > 0\ \forall i$ and $\sum_i c_i = S$, with S being a positive constant called *equivalent sample size* (ESS) [Walley, 1996]. Each θ_{x_i} can then be estimated from data as:

$$\theta_{x_i} = \frac{N[x_i] + c_i}{N + S}\,, \tag{10}$$

where $N[x_i]$ is the count of data items for which $X = x_i$, and N is the total sample size. The *imprecise* version of the Bayesian posterior estimator considers all possible Dirichlet priors for $\boldsymbol{\theta}_X$. In particular, IDM allows each c_i to vary within the interval $[0, S]$. Then, the maximum a posteriori estimate for θ_{x_i} becomes now the interval:

$$\mathcal{K}(\theta_{x_i}) = \left[\frac{N[x_i]}{N + S},\ \frac{N[x_i] + S}{N + S}\right]. \tag{11}$$

When the parent set of X is non-empty, the parameters of the conditional distributions can be derived similarly. Hence, the map f^{IDM} is:

$$f^{\mathrm{IDM}}(\theta_{x_i|\pi_X}) = \left[\frac{N[x_i, \pi_X]}{N[\pi_X] + S},\ \frac{N[x_i, \pi_X] + S}{N[\pi_X] + S}\right], \tag{12}$$

where $N[x_i, \pi_X]$ is the count of data for which $(X = x_i) \wedge (\Pi_X = \pi_X)$. Each interval is thus of size $S/(N[\pi_X] + S)$.

ε-Contamination Defense Another common approach to define a CN is known as *ε-contamination* [Cozman, 2000]. Given a distribution $P(\mathbf{X})$ and a size $\varepsilon \in (0, 1)$ for the contamination, the *ε-contamination* model considers all probability distributions of the form $R(\mathbf{X}) = (1 - \varepsilon)\, P(\mathbf{X}) + \varepsilon Q(\mathbf{X})$, where $Q(\mathbf{X})$ is any distribution on the same domain.

Given $\varepsilon \in (0, 1)$, the defense mechanism for locally and separately specified CNs is then defined as:

$$f^{\mathrm{CON}}(\theta_{x_i|\pi_X}) = \big[(1-\varepsilon)\,\theta_{x_i|\pi_X}, (1-\varepsilon)\,\theta_{x_i|\pi_X} + \varepsilon\big]. \tag{13}$$

Notice that each interval has length ε.

Random Intervals of Fixed Size Defense This approach allows us to define non-deterministic credal sets. The idea consists of *inflating* each parameter in such a way that the resulting interval: (i) includes the original point parameter; (ii) incorporates a random shift; (iii) remains closed and convex, and (iv) has a fixed size.

Let $\delta \in (0, 1]$. For any $\theta_{x_i|\pi_X}$, we can sample $\eta \sim \mathrm{Unif}(0, \delta)$ and define $\mathcal{K}(\theta_{x_i|\pi_X}) = [\theta_{x_i|\pi_X} - \eta, \theta_{x_i|\pi_X} - \eta + \delta]$. While such an interval satisfies the conditions above, its endpoints may lie outside $[0, 1]$. For this reason, additional constraints must be set. The mechanism is thus defined as follows:

$$f^{\mathrm{RAN}}(\theta_{x_i|\pi_X}) = \Big[\min\big\{1 - \delta,\ \max\{0,\ \theta_{x_i|\pi_X} - \eta\}\big\},\ \min\big\{1,\ \max\{\delta,\ \theta_{x_i|\pi_X} - \eta + \delta\}\big\}\Big],$$
$$\delta \in (0, 1], \qquad \eta \sim \mathrm{Unif}(0, \delta)\,. \tag{14}$$

It follows from Equation (14) that there are three possible intervals, depending on the values of $\theta_{x_i|\pi_X}, \eta, \delta$, and each has length δ. These intervals are $[\theta_{x_i|\pi_X} - \eta, \theta_{x_i|\pi_X} - \eta + \delta]$, $[0, \delta]$, and $[1 - \delta, 1]$.

Random Interval of Variable Size Defense This defense mechanism is a combined version of the ones above. Each interval length now matches that of f^{IDM}, but is offset by a random amount to incorporate stochastic variation. f^{LOC} is defined as in Equation (14), where δ and η get substituted by:

$$\delta_{S,\,\pi_X} = \frac{S}{N[\pi_X] + S}, \quad \eta_{S,\,\pi_X} \sim \mathrm{Unif}(0, \delta_{S,\,\pi_X})\,. \tag{15}$$

4.3 ATTACK MECHANISMS

The function g of the attack mechanism can be defined on each credal set $\mathcal{K}(\boldsymbol{\theta}_{X|\pi_X})$, for all X and π_X, so that $g(\mathcal{K}(\widehat{\boldsymbol{\theta}}_{\mathcal{T}})) = \boldsymbol{\theta}^{\mathcal{K}}_{X|\pi_X}$. In this Section, we assume that the attacker does not necessarily know which specific defense mechanism was employed before releasing the BN.

Maximum Likelihood Attack The attacker might want to be *conservative*, in the sense of seeking $\boldsymbol{\theta}^{\mathcal{K}}$ so that $L(\boldsymbol{\theta}^{\mathcal{K}} \mid \mathbf{x}) \approx L(\widehat{\boldsymbol{\theta}}_{\mathcal{R}} \mid \mathbf{x})$, $\mathbf{x} \in \mathcal{R}$. Knowing $\mathcal{R}$ and $\mathcal{G}$, taking the constrained MLE on $\mathcal{R}$ is a sensible choice:

$$g^{\mathrm{MLE}}(\mathcal{K}(\boldsymbol{\theta}_{X|\pi_X})) = \underset{\boldsymbol{\theta}^{\mathcal{K}} \in \mathcal{K}(\boldsymbol{\theta}_{X|\pi_X})}{\operatorname{argmax}} \sum_{\mathbf{x} \in \mathcal{R}} \log P_{\boldsymbol{\theta}^{\mathcal{K}}}(\mathbf{x}) \,. \quad (16)$$

Minimum Likelihood Attack In contrast to the above mechanism, the attacker might want instead to be highly *discriminative*. In other words, a BN can be sought so that the two likelihoods in the LLR function diverge. A solution is given by the BN that minimizes the likelihood on $\mathcal{R}$ or, equivalently, maximizes its negative likelihood. For this reason, g^{MNE} is defined as in Equation (16) where argmin is sought instead of argmax . Note that the optimum for g^{MNE} lies at one of the vertices of the associated credal set [Bauer, 1958], a fact that can be exploited algorithmically.

Maximum Entropy Attack In this attack, the aim is to find the BN whose distribution maximizes the entropy within the CN. Specifically:

$$g^{\mathrm{ENT}}(\mathcal{K}(\boldsymbol{\theta}_{X|\pi_X})) = \underset{\boldsymbol{\theta}^{\mathcal{K}} \in \mathcal{K}(\boldsymbol{\theta}_{X|\pi_X})}{\operatorname{argmax}} \mathrm{Ent}(P_{\boldsymbol{\theta}^{\mathcal{K}}}(X)). \quad (17)$$

where $\mathrm{Ent}(P(X)) = -\sum_{x \in \mathrm{Val}(X)} P(x) \log P(x)$ is the entropy of a distribution. The computation exploits the function's concavity.

Credal Set Centroid Attack The midpoint of each credal set (Section 4.2) can be considered an estimate of the original distribution. We adopt the center-of-mass [Miranda and Montes, 2021], so that the centroid is taken as the mean of its extreme points.

Let $\{\mathbf{v}^i_{X|\pi_X}\}_{i \in \mathcal{I}}$ be the $|\mathcal{I}|$ vertices of $\mathcal{K}(\boldsymbol{\theta}_{X|\pi_X})$. The attack is defined as:

$$g^{\mathrm{CEN}}(\mathcal{K}(\boldsymbol{\theta}_{X|\pi_X})) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathbf{v}^i_{X|\pi_X} \,. \quad (18)$$

Random Sample Attack A uniformly random mechanism g^{RAN} is considered for benchmarking purposes:

$$g^{\mathrm{RAN}}(\mathcal{K}(\boldsymbol{\theta}_{X|\pi_X})) = \boldsymbol{\theta}^{\mathcal{K}}_{X|\pi_X} \sim \mathrm{Unif}(\mathcal{K}(\boldsymbol{\theta}_{X|\pi_X})) \,. \quad (19)$$

5 EXPERIMENTS

We conduct two classes of experiments and make the source code available for research purposes [Rocchi, 2026] (please refer to the most recent version). All experiments were conducted within Docker containers to ensure environment isolation and reproducibility. The host system was equipped with two AMD EPYC 7413 processors (providing a total of 96 physical cores) and 512 GB of RAM. The exact versions of all software packages and dependencies are available in the project's repository. We note that, even for large networks, the experiments remain computationally feasible and ended within 48h.

In Section 5.1, the goal is to compare each defense against each attack mechanism with respect to the MIA power. Moreover, we investigate whether such power varies with the DAG complexity. In Section 5.2, we present experiments on the trade-off between privacy and accuracy for two models: CN and BN with Laplacian noise injection.

5.1 MODEL PRIVACY

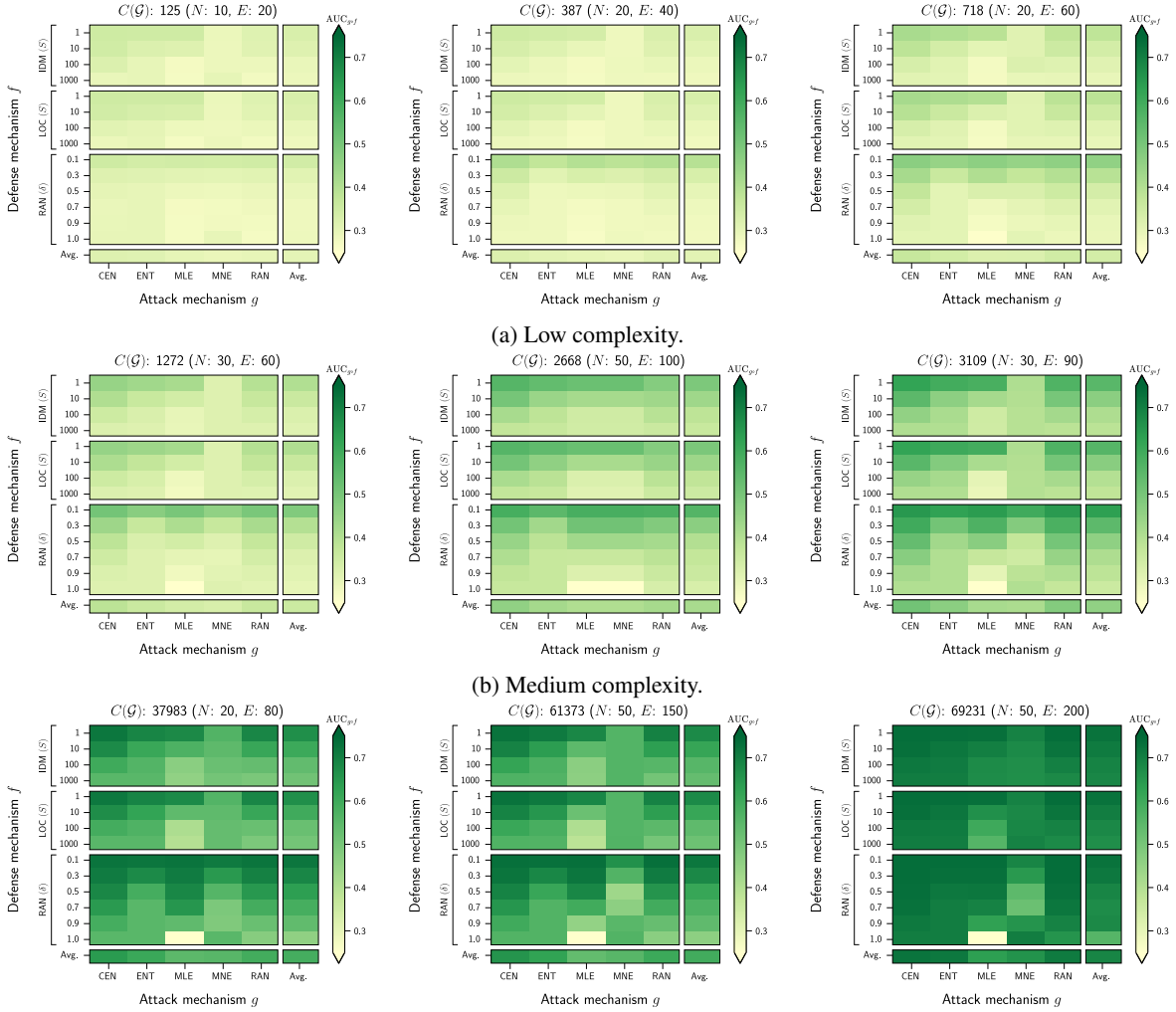
We consider all DAGs derived from the combinations of the number of nodes $M \in \{10, 20, 30, 50\}$ and edge density $E \in \{2, 3, 4\}$. The latter indicates the average number of ingoing edges per node. The consideration of large and dense networks stems from practical frameworks, such as omics data modeling [Suter et al., 2022]. Each node represents a categorical variable with at most three categories (higher-dimensional models exhibit similar behavior). The experiment l pipeline for a set of fixed (i) M nodes and $M \cdot E$ edges; (ii) defense mechanism f; (iii) attack mechanism g is given in Section B.

Figure 2 shows the information leaked (in terms of $\mathrm{AUC}_{g \circ f}(\boldsymbol{\theta}_0)$) for a subset of models, and for any pair of defense and attack mechanisms f, g. We take the area under the curve $\mathrm{AUC}_{g \circ f}(\boldsymbol{\theta}_0)$ as measure of privacy. This measure is a proxy for the information revealed by a model and may lie in the interval $[0, 1]$. A value of 1 indicates that the maximum power is achieved with the minimum error; thus, there is a complete loss of information and no privacy with respect to MIA. Conversely, a value of 0 indicates that the power over every error is null; thus, no information is revealed, and the model remains completely private.

Figure 3 shows the information leaked when the *best attack* is exploited for a pair of model $\boldsymbol{\theta}_0$ and defense f. We hereby define the *best attack* as $g^* = \operatorname{argmax}_g \mathrm{AUC}_{g \circ f}(\boldsymbol{\theta}_0)$. Overall, g^{CEN} is the best attack. Whenever it is not the case, g^* is explicitly labeled in Figure 3. We note that for $g^* \neq g^{\mathrm{CEN}}$, the difference between their respective AUCs is less than 0.005 in the majority of the cases, with the maximum obtained at $C(\mathcal{G}) = 303$ and g^{IDM} ($S = 1000$), for which the AUC difference is ≈ 0.017.

5.2 PRIVACY-ACCURACY TRADE-OFF

To evaluate the utility of a model, we assess its accuracy at a fixed privacy level using a *maximum a posteriori* (MAP) classification task. Our modeling choice is motivated by both practical and theoretical reasons. Given the privacy properties of general BNs, as examined in Section 5.1, this Section investigates the trade-off between privacy and util-



(a) Low complexity.

(b) Medium complexity.

(c) High complexity.

Figure 2: Crossed evaluation, for increasing DAG complexity. *Avg.* refers to the average of the respective row or column. Darker color means better attack, thus less privacy.

ity. For this task, we believe the CNB already shows the fundamental properties of prediction. In fact, the (credal or not credal) naive Bayes model is recognized in the literature as a remarkably robust and effective classifier [Zaffalon, 2002]. Yet, there is no significant empirical evidence that more complex models consistently outperform it on predictive tasks. Specific details applying to our framework are presented in Section D.

- In the first group of experiments, we consider a naive Bayes model with 20 nodes, from which synthetic data are generated. The target variable is binary, while the other variables have at most 4 categories. Notably, we consider 20-parameter resampling with the same model structure, so that the results do not depend on any specific distribution of the target class. As defense mechanisms, we consider f^{IDM} and f^{RAN}, while g^{CEN} is treated as the attack mechanism. Notably, f^{LOC} was excluded from the evaluation because its behavior was similar to that of the other defenses. The experimental pipeline is detailed in Section C.
- In the second group of experiments, we consider the *Breast Cancer* dataset from the *UCI Machine Learning Repository* [Zwitter and Soklic, 1988]. These data include 286 cancer patients across 1 binary target variable and 8 covariates for prediction. A CNB model is then learned from the data, where f^{IDM} and g^{CEN} were considered for the defense and attack mechanisms, respectively.

For each credal model considered, along with its corresponding MIA power, we seek a noisy model that is equally private to compare with. We consider the model obtained by injecting Laplacian noise into $\widehat{\boldsymbol{\theta}}_{\mathcal{T}}$, following the *PrivBayes* approach described in Zhang et al. [2017]. The hyperparameter ϵ controls the level of differential privacy provided

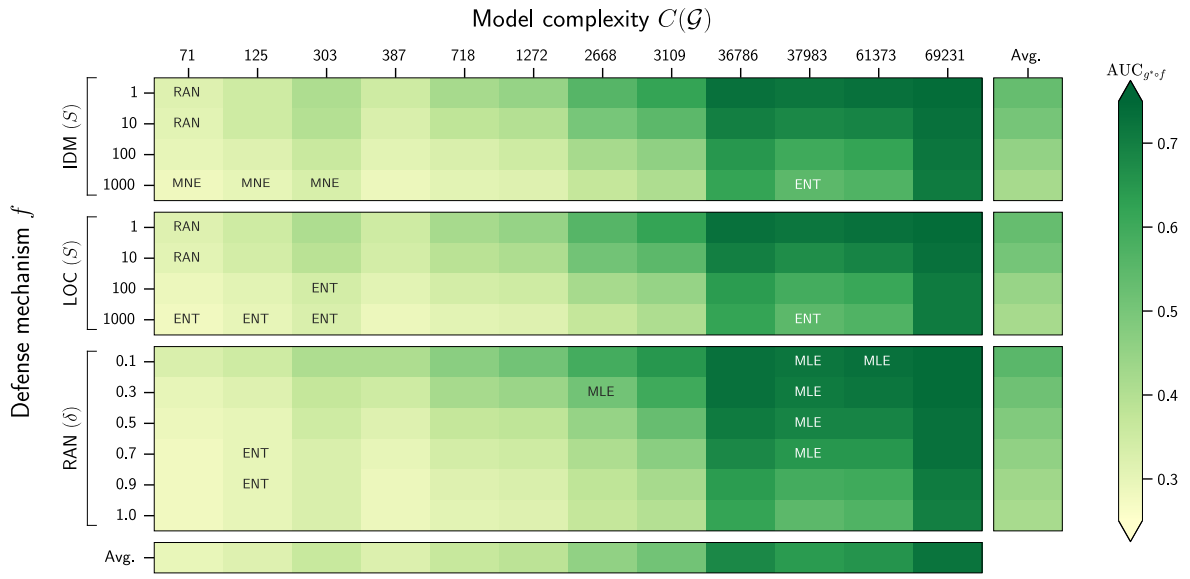
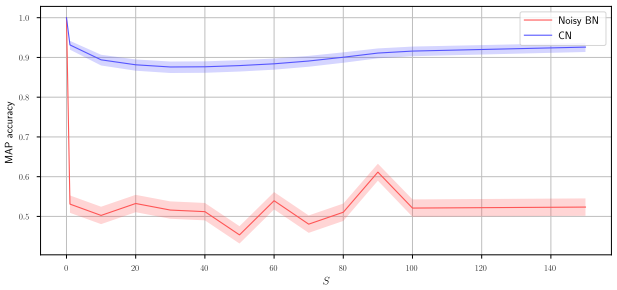


Figure 3: Best attack against each defense, as DAG complexity increases. *Avg.* refers to the average of the respective row or column. g^{CEN} is the best attack for the unnamed cells.

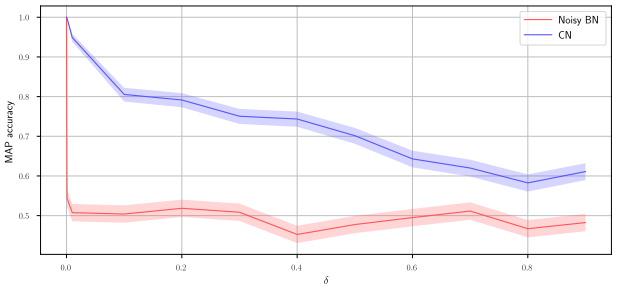
by the noisy model. While we acknowledge other noise mechanisms that guarantee ϵ-DP, we note that our choice of the Laplacian distribution is an intermediate step toward achieving a privacy-guaranteed model (at a fixed ϵ).

In particular, ϵ is related to the scale parameter λ of the Laplacian distribution from which the noise is sampled. Given ϵ, and by setting $\lambda = \frac{2M}{|\mathcal{P}|\epsilon}$, the model satisfies the ϵ-DP [Zhang et al., 2017]. For a given S (or δ), we compute the value of ϵ so that the power-vs-error curves of the two attacked models are *similar*. Specifically, we choose ϵ so that the absolute difference between the two AUCs is within the 1% tolerance. Notice that the AUC is computed for Type I errors $\alpha \leqslant 0.1$, as detailed in Section C and supported by the literature [Carlini et al., 2022]. Larger values of α exhibit similar behavior.

Figure 4a describes the results for f^{IDM} and g^{CEN} for the synthetic data. The figure depicts the MAP accuracies for the CN and noisy BN models. The CN predictions are not divided into *certain* and *uncertain* (see Section D). The reason is that only $S = 1$ yields a nonzero ratio for certain predictions; hence, all predictions for $S > 1$ are uncertain. The values of ϵ related to different S are all close to 10^{-8}. This value of ϵ is so low that the noisy BN predictions are almost all random, i.e., the noisy BN achieves accuracies around 0.5. For higher privacy requirements, the credal model seems to achieve a clear superiority in accuracy. However, we believe that the increasing behavior related to the CN is due to numerical instability rather than a definitive trend. In fact, it might resemble a plateau or random peak that does not persist at larger values of S. This may depend on several factors, such as the experiment's hyperparameters.



(a) f^{IDM} vs g^{CEN}.



(b) f^{RAN} vs g^{CEN}.

Figure 4: MAP accuracy results on synthetic naive Bayes models and data. Hues indicate Wilson 95% confidence intervals.

Similar results are obtained for f^{RAN} and g^{CEN}, which are portrayed in Figure 4b. In this case, all CN predictions are uncertain for $\delta > 0.1$; even for $\delta = 0.1$, the ratio of certain predictions is lower than 10%. All values of ϵ are around 10^{-8}; however, we see the CN accuracy degradates as δ increases, as opposed to Figure 4a. Yet, the credal model

shows superior performance to the noisy one.

Similar results are obtained with the Breast Cancer dataset, as depicted in Figure 5.

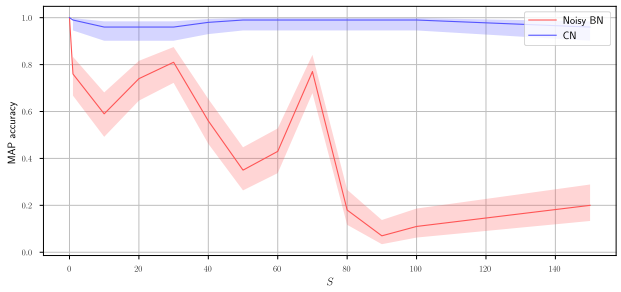


Figure 5: MAP accuracy results on real-world data with f^{IDM} and g^{CEN}. Hues indicate Wilson 95% confidence intervals.

6 CONCLUSIONS

This work explores how the imprecise probability framework can effectively balance privacy and accuracy in BNs, providing new evidence to existing research. First, we formalize an imprecise MIA framework through the interaction between defense mechanisms that map precise BN parameters to CN parameter sets and attack mechanisms that select representative parameters from those sets. Then, we present empirical results demonstrating how and when uncertainty-based defense mechanisms reduce attack effectiveness, particularly as uncertainty in the CN increases. While there is no universal solution for defending the models considered, the centroid-based attack emerges as the most powerful option for adversaries. Additionally, experiments on the CNB classification reveal that credal models can achieve competitive or superior predictive performance compared to differentially private models at similar privacy levels. Our focus on locally and separately specified CNs is a specific choice. Yet, we believe this class of models already shows most of the privacy-related properties of more sophisticated models.

Future research may investigate different versions of the proposed MIA framework, such as: a credal version of the LLR function, where the attacker exploits probability bounds instead of point estimates; adaptive attackers that have access to more information, or that can query the model multiple times; novel defenses that make the centroid attack less effective or computationally unfeasible. In particular, an improved test statistic can be constructed, inspired by the KL decomposability in Section A; that is, an attacker could leverage rare configurations of the BN to identify specific individuals rather than adopting the (global) LLR function. A similar argument applies to the PrivBayes approach [Zhang et al., 2017], where the amount of noise injected does not account for the amount of data available to train each parameter. Thus, our main competitor has also room for improving their privacy-accuracy trade-off.

Besides, it is worthwhile to consider the model's utility measures beyond classification accuracy. Moreover, comprehensive methodical analysis of the relationship between credal models and differential privacy could guide privacy research to prioritize uncertain models for privacy preservation rather than relying solely on noise injection.

Author Contributions

Niccolò Rocchi: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Fabio Stella**: Funding acquisition, Project administration, Resources, Supervision, Validation, Writing – review & editing. **Cassio de Campos**: Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Acknowledgements

This work was supported by the MUR under the grant "Dipartimenti di Eccellenza 2023-2027" of the Department of Informatics, Systems and Communication of the University of Milano-Bicocca, Milan, Italy, and by the National Plan for NRRP Complementary Investments (Project n. PNC0000003 - AdvaNced Technologies for Human-centrEd Medicine (ANTHEM)). Niccolò Rocchi is funded by NextGeneration EU and Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy, in the Horizon Europe project IDEA4RC framework. Cassio de Campos thanks the support from the Dutch Research Council (NWO) via project NGF.1609.242.024.

References

Alessandro Antonucci, Cassio P. de Campos, David Huber, and Marco Zaffalon. Approximate credal network updating by linear programming with applications to decision making. *International Journal of Approximate Reasoning*, 58:25–38, March 2015. ISSN 0888-613X. doi: 10.1016/j.ijar.2014.10.003.

Heinz Bauer. Minimalstellen von funktionen und extremalpunkte. *Archiv der Mathematik*, 9(4):389–393, November 1958. ISSN 1420-8938. doi: 10.1007/bf01898615.

Ruta Binkyte, Carlos Antonio Pinzón, Szilvia Lestyán, Kangsoo Jung, Héber Hwang Arcolezi, and Catuscia

Palamidessi. Causal discovery under local privacy. In Francesco Locatello and Vanessa Didelez, editors, *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 325–383. PMLR, 2024.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE, May 2022. doi: 10.1109/sp46214.2022.9833649.

Anthony C. Constantinou, Zhigao Guo, and Neville K. Kitson. The impact of prior knowledge on causal structure learning. *Knowledge and Information Systems*, 65(8):3385–3434, 2023. ISSN 0219-3116. doi: 10.1007/s10115-023-01858-x.

G. Corani and C.P. de Campos. A tree augmented classifier based on extreme imprecise dirichlet model. *International Journal of Approximate Reasoning*, 51(9):1053–1068, November 2010. ISSN 0888-613X. doi: 10.1016/j.ijar.2010.08.007.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, October 2001. ISBN 9780471200611. doi: 10.1002/0471200611.

Fabio G. Cozman. Credal networks. *Artificial Intelligence*, 120(2):199–233, 2000. ISSN 0004-3702. doi: 10.1016/s0004-3702(00)00029-1.

Fabio Gagliardi Cozman and Cassio P. de Campos. *Local computation in credal networks*, pages 5–11. ECAI, 2004.

Anirban DasGupta. *Asymptotic theory of statistics and probability*. SpringerLink. Springer Science+Business Media, LLC, New York, NY, 2008. ISBN 9780387759715. Includes bibliographical references and index.

C. P. de Campos and F. G. Cozman. *Inference in Credal Networks Through Integer Programming*, pages 145–154. SIPTA, 2007.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. *Journal of Privacy and Confidentiality*, 7(3):17–51, 2017. ISSN 2575-8527. doi: 10.29012/jpc.v7i3.405.

Marcel Grube, Casper Reijnen, Peter J. F. Lucas, Frieder Kommoss, Felix K. F. Kommoss, Sara Y. Brucker, Christina B. Walter, Ernst Oberlechner, Bernhard Krämer, Jürgen Andress, Felix Neis, Annette Staebler, Johanna M. A. Pijnenborg, and Stefan Kommoss. Improved preoperative risk stratification in endometrial carcinoma patients: external validation of the endorisk bayesian network model in a large population-based case series. *Journal of Cancer Research and Clinical Oncology*, 149(7):3361–3369, August 2022. doi: 10.1007/s00432-022-04218-4.

Daphne Koller and Nir Friedman. *Probabilistic graphical models*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass. [u.a.], [nachdr.] edition, 2010. ISBN 9780262013192.

Sasi Kumar Murakonda, Reza Shokri, and George Theodorakopoulos. Quantifying the privacy risks of learning high-dimensional graphical models. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2287–2295. PMLR, 2021.

E.L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer International Publishing, 2022. ISBN 9783030705787. doi: 10.1007/978-3-030-70578-7.

Enrique Miranda and Ignacio Montes. *Centroids of Credal Sets: A Comparative Study*, pages 427–441. Springer International Publishing, 2021. ISBN 9783030867720. doi: 10.1007/978-3-030-86772-0_31.

Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694–706):289–337, February 1933. ISSN 2053-9258. doi: 10.1098/rsta.1933.0009.

Judea Pearl. *From Bayesian Networks to Causal Networks*, pages 157–182. Springer US, 1995. ISBN 9781489914248. doi: 10.1007/978-1-4899-1424-8_9.

Niccolò Rocchi. Niccolo-rocchi/bn-privacy: Imprecise probabilities for privacy–accuracy trade-offs in bayesian networks, June 2026. URL `https://doi.org/10.5281/zenodo.20760736`.

Niccolò Rocchi, Fabio Stella, and Cassio de Campos. *Towards Privacy-Aware Bayesian Networks: A Credal Approach*. IOS Press, October 2025. ISBN 9781643686318. doi: 10.3233/faia251419.

Niccolò Rocchi, Alessio Zanga, Alice Bernasconi, Alessandro Gronchi, Dario Callegaro, Alessandra Borghi, Paolo Giovanni Casali, Salvatore Provenzano, Rosalba Miceli, Annalisa Trama, and Fabio Stella. A causal discovery workflow for rare diseases: Experts-in-the-loop analysis of sparse longitudinal data. *Journal of Medical Systems*, 50(1), January 2026. ISSN 1573-689X. doi: 10.1007/s10916-025-02327-4.

Saptarshi Roy, Raymond K. W. Wong, and Yang Ni. Directed cyclic graph for causal discovery from multivariate functional data. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 42762–42774. Curran Associates, Inc., 2023.

Mauro Scanagatta, Antonio Salmerón, and Fabio Stella. A survey on bayesian network structure learning from data. *Progress in Artificial Intelligence*, 8(4):425–439, 2019. ISSN 2192-6360. doi: 10.1007/s13748-019-00194-y.

Polina Suter, Eva Dazert, Jack Kuipers, Charlotte K. Y. Ng, Tuyana Boldanova, Michael N. Hall, Markus H. Heim, and Niko Beerenwinkel. Multi-omics subtyping of hepatocellular carcinoma patients using a bayesian network mixture model. *PLOS Computational Biology*, 18(9):e1009767, 2022. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1009767.

Pablo Torrijos, José A. Gámez, and José M. Puerta. *FedGES: A Federated Learning Approach for Bayesian Network Structure Learning*, pages 83–98. Springer Nature Switzerland, 2025. ISBN 9783031789809. doi: 10.1007/978-3-031-78980-9_6.

Peter Walley. Inferences from multinomial data: Learning about a bag of marbles. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):3–34, 1996. ISSN 1467-9868. doi: 10.1111/j.2517-6161.1996.tb02065.x.

Peter Walley. Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24(2–3):125–148, May 2000. ISSN 0888-613X. doi: 10.1016/s0888-613x(00)00031-1.

S. S. Wilks. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938. doi: 10.1214/aoms/1177732360.

Marco Zaffalon. The naive credal classifier. *Journal of Statistical Planning and Inference*, 105(1):5–21, 2002. ISSN 0378-3758. doi: 10.1016/s0378-3758(01)00201-4.

Alessio Zanga, Alice Bernasconi, Peter J. F. Lucas, Hanny Pijnenborg, Casper Reijnen, Marco Scutari, and Fabio Stella. Causal discovery with missing data in a multicentric clinical study. In Jose M. Juarez, Mar Marcos, Gregor Stiglic, and Allan Tucker, editors, *Artificial Intelligence in Medicine*, pages 40–44, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-34344-5.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024. doi: 10.5555/3692070.3694473.

Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. *ACM Transactions on Database Systems*, 42(4):1–41, 2017. doi: 10.1145/3134428.

Matjaz Zwitter and Milan Soklic. Breast Cancer. UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C51P4M.

Imprecise Probabilities for Privacy–Accuracy Trade-offs in Bayesian Networks (Supplementary Material)

Niccolò Rocchi[1,2] **Fabio Stella**[1] **Cassio de Campos**[3]

[1]University of Milano-Bicocca, Milan, Italy
[2]Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[3]Eindhoven University of Technology, Eindhoven, The Netherlands

A THEORETICAL ANALYSIS

In this section, we exploit well-known results from information theory [Cover and Thomas, 2001] and asymptotic theory [Lehmann and Romano, 2022, DasGupta, 2008].

Theorem A.1 (Asymptotic statistic under H_0)**.** *Let $\{\mathbf{x}_i\}_{i=1}^n$ a set of random instances of $\mathbf{X}$ and assume that H_0 holds, namely, $P(\mathbf{X};\boldsymbol{\theta}_0)$ is the distribution from which all $\mathbf{x}_i$ have been generated. Hence, we have:*

$$\Lambda(\{\mathbf{x}_i\}_{i=1}^n, \boldsymbol{\theta}^{\mathcal{K}}) \xrightarrow{d} \mathcal{N}(-\mu_0,\, 2\mu_0) \quad \forall \boldsymbol{\theta}^{\mathcal{K}}\,, \tag{20}$$

where $\mu_0 = \mathrm{KL}(\widehat{\boldsymbol{\theta}}_{\mathcal{R}} \,\|\, \boldsymbol{\theta}^{\mathcal{K}})$ is the Kullback-Leibler (KL) divergence.

Proof. Assume H_0 holds, and that $\widehat{\boldsymbol{\theta}}_{\mathcal{R}}$ is the MLE for $\boldsymbol{\theta}_0$, as stated in Section 3. Hereafter, we drop $\{\mathbf{x}_i\}_{i=1}^n$ in all subsequent formulas for notational convenience. We recall that $\Lambda(\boldsymbol{\theta}^{\mathcal{K}}) = L(\boldsymbol{\theta}^{\mathcal{K}}) - L(\widehat{\boldsymbol{\theta}}_{\mathcal{R}})$ for $\boldsymbol{\theta}^{\mathcal{K}} \in \mathcal{K}(\widehat{\boldsymbol{\theta}}_{\mathcal{T}})$, as defined in Eq. (9).

Let $\boldsymbol{\theta}^{\mathcal{K}}, \widehat{\boldsymbol{\theta}}_{\mathcal{R}} \in [0,1]$ be point parameters. Using the Taylor expansion, we can write:

$$L(\boldsymbol{\theta}^{\mathcal{K}}) \;\approx\; L(\widehat{\boldsymbol{\theta}}_{\mathcal{R}}) + (\boldsymbol{\theta}^{\mathcal{K}} - \widehat{\boldsymbol{\theta}}_{\mathcal{R}})L'(\widehat{\boldsymbol{\theta}}_{\mathcal{R}}) + \frac{1}{2}(\boldsymbol{\theta}^{\mathcal{K}} - \widehat{\boldsymbol{\theta}}_{\mathcal{R}})^2 L''(\widehat{\boldsymbol{\theta}}_{\mathcal{R}})\,. \tag{21}$$

Here $L'(\cdot)$ and $L''(\cdot)$ are the first and second derivatives of $L(\cdot)$. Because of the central limit theorem, the first derivative $L'(\widehat{\boldsymbol{\theta}}_{\mathcal{R}})$ converges in distribution to a Normal; its mean is $\mathbb{E}[L'(\widehat{\boldsymbol{\theta}}_{\mathcal{R}})] = 0$, due to the MLE definition, and its variance is $\mathrm{Var}[L'(\widehat{\boldsymbol{\theta}}_{\mathcal{R}})] = \mathcal{I}(\widehat{\boldsymbol{\theta}}_{\mathcal{R}})$, where $\mathcal{I}(\widehat{\boldsymbol{\theta}}_{\mathcal{R}})$ is the Fisher information. Hence, we have:

$$L'(\widehat{\boldsymbol{\theta}}_{\mathcal{R}}) \xrightarrow{d} \mathcal{N}(0,\, \mathcal{I}(\widehat{\boldsymbol{\theta}}_{\mathcal{R}}))\,. \tag{22}$$

Using the weak law of large numbers, we also know that the second derivative converges in probability:

$$L''(\widehat{\boldsymbol{\theta}}_{\mathcal{R}}) \xrightarrow{p} -\mathcal{I}(\widehat{\boldsymbol{\theta}}_{\mathcal{R}})\,. \tag{23}$$

By exploiting the asymptotic results in Equation (21) and by invoking Slutsky's theorem, we derive the following convergence in distribution:

$$\Lambda(\boldsymbol{\theta}^{\mathcal{K}}) \;\approx\; (\boldsymbol{\theta}^{\mathcal{K}} - \widehat{\boldsymbol{\theta}}_{\mathcal{R}})\, L'(\widehat{\boldsymbol{\theta}}_{\mathcal{R}}) + \frac{1}{2}\,(\boldsymbol{\theta}^{\mathcal{K}} - \widehat{\boldsymbol{\theta}}_{\mathcal{R}})^2\, L''(\widehat{\boldsymbol{\theta}}_{\mathcal{R}}) \xrightarrow{d} \mathcal{N}(-\mu_0, 2\mu_0)\,, \tag{24}$$

where $\mu_0 = \frac{1}{2}\,(\boldsymbol{\theta}^{\mathcal{K}} - \widehat{\boldsymbol{\theta}}_{\mathcal{R}})^2\, \mathcal{I}(\widehat{\boldsymbol{\theta}}_{\mathcal{R}})$. We recall that the derivatives of the KL with respect to the first argument depend on the Fisher information. For this reason:

$$\mu_0 = \frac{1}{2}\,(\boldsymbol{\theta}^{\mathcal{K}} - \widehat{\boldsymbol{\theta}}_{\mathcal{R}})^2\, \mathcal{I}(\widehat{\boldsymbol{\theta}}_{\mathcal{R}}) = \mathrm{KL}(\widehat{\boldsymbol{\theta}}_{\mathcal{R}} \,\|\, \boldsymbol{\theta}^{\mathcal{K}})\,. \tag{25}$$

Similar arguments follow when $\widehat{\boldsymbol{\theta}}_{\mathcal{R}}$ and $\boldsymbol{\theta}^{\mathcal{K}}$ are vectors representing multivariate distributions. In that case, we have:

$$\Lambda(\boldsymbol{\theta}^{\mathcal{K}}) \;\approx\; (\boldsymbol{\theta}^{\mathcal{K}} - \widehat{\boldsymbol{\theta}}_{\mathcal{R}})^T \nabla_\theta L(\widehat{\boldsymbol{\theta}}_{\mathcal{R}}) + \frac{1}{2} (\boldsymbol{\theta}^{\mathcal{K}} - \widehat{\boldsymbol{\theta}}_{\mathcal{R}})^T \nabla^2_{\theta\theta} L(\widehat{\boldsymbol{\theta}}_{\mathcal{R}}) (\boldsymbol{\theta}^{\mathcal{K}} - \widehat{\boldsymbol{\theta}}_{\mathcal{R}}) \xrightarrow{d} \mathcal{N}(-\mu_0, 2\mu_0)\,, \tag{26}$$

where $\mu_0 = \frac{1}{2} (\boldsymbol{\theta}^{\mathcal{K}} - \widehat{\boldsymbol{\theta}}_{\mathcal{R}})^T \mathcal{I}(\widehat{\boldsymbol{\theta}}_{\mathcal{R}}) (\boldsymbol{\theta}^{\mathcal{K}} - \widehat{\boldsymbol{\theta}}_{\mathcal{R}}) = \mathrm{KL}(\widehat{\boldsymbol{\theta}}_{\mathcal{R}} \,\|\, \boldsymbol{\theta}^{\mathcal{K}})$. Here, $\mathcal{I}(\widehat{\boldsymbol{\theta}}_{\mathcal{R}})$ is the Fisher information matrix. □

Theorem A.2 (Asymptotic statistic under H_1)**.** *Let $\{\mathbf{x}_i\}_{i=1}^n$ a set of random instances of $\mathbf{X}$ and assume that H_1 holds, namely, $P(\mathbf{X}; \widehat{\boldsymbol{\theta}}_{\mathcal{T}})$ is the distribution from which all $\mathbf{x}_i$ have been generated. Moreover, assume that $\boldsymbol{\theta}^{\mathcal{K}} \to \widehat{\boldsymbol{\theta}}_{\mathcal{T}}$ for $|\mathcal{T}| \to \infty$. Hence, we have:*

$$\Lambda(\{\mathbf{x}_i\}_{i=1}^n, \boldsymbol{\theta}^{\mathcal{K}}) \xrightarrow{d} \mathcal{N}(\mu_1,\, 2\mu_1)\,, \tag{27}$$

where $\mu_1 = \mathrm{KL}(\boldsymbol{\theta}^{\mathcal{K}} \,\|\, \widehat{\boldsymbol{\theta}}_{\mathcal{R}})$.

Proof. Analogous to the proof of Theorem A.1 with the positions of $\widehat{\boldsymbol{\theta}}_{\mathcal{R}}$ and $\boldsymbol{\theta}^{\mathcal{K}}$ interchanged. □

Note that the difference between the means of the distributions under H_0 and H_1 is the *Jeffreys divergence* between $\widehat{\boldsymbol{\theta}}_{\mathcal{R}}$ and $\boldsymbol{\theta}^{\mathcal{K}}$, which depends on the model complexity. Specifically, each KL factorizes into a weighted sum of *local* KLs according to the graph. By leveraging the definition of KL and the BN factorization in Eq. (1), we obtain:

$$\begin{aligned}
\mathrm{KL}(\widehat{\boldsymbol{\theta}}_{\mathcal{R}} \,\|\, \boldsymbol{\theta}^{\mathcal{K}}) \;&=\; \sum_{\mathbf{x} \in \mathrm{Val}(\mathbf{X})} P_{\widehat{\boldsymbol{\theta}}_{\mathcal{R}}}(\mathbf{x}) \,\log \frac{P_{\widehat{\boldsymbol{\theta}}_{\mathcal{R}}}(\mathbf{x})}{P_{\boldsymbol{\theta}^{\mathcal{K}}}(\mathbf{x})} && (28)\\
&= \sum_{\mathbf{x} \in \mathrm{Val}(\mathbf{X})} P_{\widehat{\boldsymbol{\theta}}_{\mathcal{R}}}(\mathbf{x}) \sum_{i=1}^{m} \log \frac{P_{\widehat{\boldsymbol{\theta}}_{\mathcal{R}}}(x_i \mid \pi_i)}{P_{\boldsymbol{\theta}^{\mathcal{K}}}(x_i \mid \pi_i)} && (29)\\
&= \sum_{i=1}^{m} \sum_{\substack{x_i \in \\ \mathrm{Val}(X_i)}} \sum_{\substack{\pi_i \in \\ \mathrm{Val}(\Pi_{X_i})}} \sum_{\substack{\mathbf{x} \in \mathrm{Val}(\mathbf{X}) \\ \setminus \{X_i, \Pi_i\}}} P_{\widehat{\boldsymbol{\theta}}_{\mathcal{R}}}(\mathbf{x}) \,\log \frac{P_{\widehat{\boldsymbol{\theta}}_{\mathcal{R}}}(x_i \mid \pi_i)}{P_{\boldsymbol{\theta}^{\mathcal{K}}}(x_i \mid \pi_i)} && (30)\\
&= \sum_{i=1}^{m} \sum_{\substack{x_i \in \\ \mathrm{Val}(X_i)}} \sum_{\substack{\pi_i \in \\ \mathrm{Val}(\Pi_{X_i})}} \log \frac{P_{\widehat{\boldsymbol{\theta}}_{\mathcal{R}}}(x_i \mid \pi_i)}{P_{\boldsymbol{\theta}^{\mathcal{K}}}(x_i \mid \pi_i)} \, P_{\widehat{\boldsymbol{\theta}}_{\mathcal{R}}}(x_i \mid \pi_i) \, P_{\widehat{\boldsymbol{\theta}}_{\mathcal{R}}}(\pi_i)\,. && (31)
\end{aligned}$$

The term $\log \frac{P_{\widehat{\boldsymbol{\theta}}_{\mathcal{R}}}(x_i|\pi_i)}{P_{\boldsymbol{\theta}^{\mathcal{K}}}(x_i|\pi_i)} \, P_{\widehat{\boldsymbol{\theta}}_{\mathcal{R}}}(x_i \mid \pi_i)$ inside the summations matches the definition of KL for two parameters explaining $P(x_i, \mid \pi_i)$. Consequently, complex models tend to be more susceptible to MIA attacks.

Let now α be a fixed Type-I error, and z_α the quantile at level $1-\alpha$ of the standard Normal. Under H_0, the critical threshold for the test is $\tau(\alpha) = -\mathrm{KL}(\widehat{\boldsymbol{\theta}}_{\mathcal{R}} \| \boldsymbol{\theta}^{\mathcal{K}}) + z_\alpha \sqrt{2\mathrm{KL}(\widehat{\boldsymbol{\theta}}_{\mathcal{R}} \| \boldsymbol{\theta}^{\mathcal{K}})}$. The function $f : \mathbb{R}^+ \to \mathbb{R}, \quad f(x) = -x + z_\alpha \sqrt{2x}$ is concave and the point of global maximum is $\frac{z_\alpha^2}{2}$. For $x \geqslant \frac{z_\alpha^2}{2}$, f is strictly decreasing. This gives a relation between the threshold and the divergency between the model $\widehat{\boldsymbol{\theta}}_{\mathcal{R}}$ trained with reference population and the choice of attack $\boldsymbol{\theta}^{\mathcal{K}}$.

B EXPERIMENTAL SETUP FOR MODEL PRIVACY

1. **Data generation.** A DAG $\mathcal{G}$ over M nodes and $M \cdot E$ edges is randomly generated. The set of parameters $\boldsymbol{\theta}_0$ is also randomly generated, based on the DAG structure. The BN $(\mathcal{G}, \boldsymbol{\theta}_0)$ is taken as the ground truth for the data generation process. Then, we sample 5 000 i.i.d. data records from $(\mathcal{G}, \boldsymbol{\theta}_0)$ and consider them as the general population $\mathcal{P}$. Two sets $\mathcal{R}, \mathcal{T}$ are randomly sampled from $\mathcal{P}$ with a ratio of 0.5 and 0.25, respectively. Next, $\widehat{\boldsymbol{\theta}}_{\mathcal{T}}$ and $\widehat{\boldsymbol{\theta}}_{\mathcal{R}}$ are estimated through MLE from the respective data.
2. **Defense & attack mechanisms.** The learned $\widehat{\boldsymbol{\theta}}_{\mathcal{T}}$ is protected by estimating a CN through the chosen f. The hyperparameter S is chosen for f^{IDM} and f^{LOC}, while δ is chosen for f^{RAN}. We considered $S = \{1, 10, 100, 1000\}$ and $\delta = \{0.1, 0.3, \ldots, 0.9, 1.0\}$ throughout the experiments. A simulated attacker, having access to the CN $f(\widehat{\boldsymbol{\theta}}_{\mathcal{T}})$, exploits g to select a BN. Once $(\mathcal{G}, \boldsymbol{\theta}^{\mathcal{K}})$ is obtained, the attacker moves to the next phase.

3. **Membership inference.** The MIA attack is conducted on both models, $\widehat{\boldsymbol{\theta}}_{\mathcal{T}}$ and $\boldsymbol{\theta}^{\mathcal{K}}$, using the test statistics $\Lambda(\cdot)$ and $\Lambda(\cdot, \boldsymbol{\theta}^{\mathcal{K}})$, respectively. Namely, given a sequence of Type I errors $\boldsymbol{\alpha} = \alpha_1, \ldots, \alpha_k$, log-spaced in $[10e^{-4}, 1)$, $k = 30$, the aim of this phase is to compute the corresponding power, i.e, $\boldsymbol{\beta} = \beta_1, \ldots, \beta_k$ for each model. For any α_i, we consider $\mathbf{x} \in \mathcal{P} \setminus \mathcal{R}$ to derive β_i. In particular, the test related to MIA assigns either $\mathbf{x} \notin \mathcal{T}$ (H_0) or $\mathbf{x} \in \mathcal{T}$ (H_1). The true positive rate β_i is computed for any error α_i. For benchmarking, we also compute the theoretical power estimate given by Theorem 3.1, with $C(\mathcal{G})$ denoting the complexity of $\mathcal{G}$ and $|\mathcal{T}|$ the target sample size.

4. **Evaluation metric.** To ensure variability in the sample distributions of $\mathcal{R}$ and $\mathcal{T}$ in step 1, these two are subsampled 200 times from $\mathcal{P}$, resulting in pairs $\{(\mathcal{R}_j, \mathcal{T}_j) : j = 1, \ldots, 200\}$. Hence, the above steps (1-3) are repeated for each subsample, yielding $\boldsymbol{\beta}_j$ for $j = 1, \ldots, 200$. Finally, the power $\boldsymbol{\beta}$ of MIA against the model $(\mathcal{G}, \boldsymbol{\theta}_0)$ and associated with $g \circ f$ is taken as the average of $\boldsymbol{\beta}_j{}'s$. The pair $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ defines a curve which is strictly related to privacy or, in other terms, how much information is leaked. As an evaluation metric for the *privacy* of a given model, related to $g \circ f$, we take the related area under the curve $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. As a notation, we denote it as $\mathrm{AUC}_{g \circ f}(\boldsymbol{\theta}_0)$. This measure is a proxy for the information revealed by a model and may lie in the interval $[0, 1]$. A value of 1 indicates that the maximum power is achieved with the minimum error; thus, there is a complete loss of information and no privacy with respect to MIA. Conversely, a value of 0 indicates that the power over every error is null; thus, no information is revealed, and the model remains completely private.

C EXPERIMENTAL SETUP FOR PRIVACY-ACCURACY TRADE-OFF

The *credal naive Bayes* (CNB) model [Zaffalon, 2002] shares the same structure as the standard naive Bayes model: we have (i) a target variable T and a set of covariates $X_1, \ldots, X_{M-1}$; (ii) edges $T \to X_i$, for any i.

1. **Data generation.** Once the model's structure $\mathcal{G}$ has been set, the parameter set $\boldsymbol{\theta}_0$ is randomly generated based on the DAG structure. Hence, $(\mathcal{G}, \boldsymbol{\theta}_0)$ is taken as the ground truth for the data generation process. Thus, we sample 1 000 i.i.d. data records from $(\mathcal{G}, \boldsymbol{\theta}_0)$ and consider them as the general population $\mathcal{P}$. Two sets $\mathcal{R}, \mathcal{T}$ are randomly sampled from $\mathcal{P}$ with a ratio of 0.5 and 0.25, respectively. Next, $\widehat{\boldsymbol{\theta}}_{\mathcal{T}}$ and $\widehat{\boldsymbol{\theta}}_{\mathcal{R}}$ are estimated through MLE from the respective data.

2. **Mechanisms & membership inference.** The defense and attack mechanisms, as well as the following MIA, are performed as described in steps 2-3 of Section B. The only difference consists in the $\boldsymbol{\alpha} = \alpha_1, \ldots, \alpha_k$ considered, which are now log-spaced in $[10e^{-4}, 0.1]$, $k = 30$ [Carlini et al., 2022]. Furthermore, we consider $S = \{1, 10, 20, \ldots, 100, 150\}$ for f^{IDM} and $\delta = \{0.001, 0.01, 0.1, 0.2, \ldots, 0.9\}$ for f^{RAN}.

3. **Comparing accuracy.** This step compares the accuracy of the two models (CN and the noisy one). To evaluate the accuracy of a given model for a fixed level of privacy, we conduct a *maximum a posteriori* (MAP) classification task. We run 100 inferences for each model, by randomly sampling $\mathbf{x} \in \mathrm{Val}(T, X_1, \ldots, X_{M-1})$. Steps 1-4 are reiterated to enforce the generalizability of results. We consider 20 random parametrizations of the ground-truth model in step 1, to ensure sufficient parametric variability. Moreover, in the same step 1, the subsampling of $\mathcal{R}, \mathcal{T}$ from $\mathcal{P}$ is performed 50 times, similarly to Section 5.1. Hence, the accuracy of each model is averaged among all these repetitions.

D MAP ESTIMATION IN THE CREDAL NAIVE BAYES MODEL

Derivations here follow from [Zaffalon, 2002]. We consider an NB model with a target variable T and a set of covariates $X_1, \ldots, X_{M-1}$. In the BN model, the target variable T is the parent of any other variable, forming a star-like structure $X_1, \ldots, X_{M-1}$; (ii) edges $T \to X_i$, for any i. If θ is the set of parameters associated with the model, it follows from Eq. 1:

$$P_\theta(T, \mathbf{X}) = P_{\theta_T}(T) \prod_{i=1}^{M-1} P_{\theta_i}(X \mid T)\,, \tag{32}$$

where the θ_i's parameterized the local distributions.

We focus on T being binary, that is, $t \in \{0, 1\}$. In the following, we provide a computationally efficient derivation of the MAP estimation for the credal NB model. Hereafter, we omit specifying the parameters in the probabilities and distributions.

$$P(t \mid \mathbf{x}) = \frac{P(t, \mathbf{x})}{P(\mathbf{x})} = \frac{P(t, \mathbf{x})}{P(t, \mathbf{x}) + P(t', \mathbf{x})}\,, \tag{33}$$

where $t' = \neg t$. As we consider locally and separately specified CNs, it holds:

$$\min_{\mathcal{K}(\mathbf{X})} P(t \mid \mathbf{x}) = \underline{P}(t \mid \mathbf{x}) = \frac{\underline{P}(t, \mathbf{x})}{\underline{P}(t, \mathbf{x}) + \overline{P}(t', \mathbf{x})}\,. \tag{34}$$

We now exploit the logarithm for numerical stability:

$$P(t, \mathbf{x}) = P(t) \cdot \prod_{i=1}^{n} P(x_i \mid t), \tag{35}$$

$$\log P(t, \mathbf{x}) = \log P(t) + \sum_{i=1}^{n} \log P(x_i \mid t), \tag{36}$$

$$\log \underline{P}(t, \mathbf{x}) = \log \underline{P}(t) + \sum_{i=1}^{n} \log \underline{P}(x_i \mid t), \tag{37}$$

$$\log \overline{P}(t', \mathbf{x}) = \log \overline{P}(t') + \sum_{i=1}^{n} \log \overline{P}(x_i \mid t')\,. \tag{38}$$

We set the notation $l\underline{p}_t := \log \underline{P}(t, \mathbf{x})$. Similarly, we define $l\overline{p}_t, l\underline{p}_{t'}, l\overline{p}_{t'}$. We obtain:

$$\log \underline{P}(t \mid \mathbf{x}) = l\underline{p}_t - \log[\exp(l\underline{p}_t) + \exp(l\overline{p}_{t'})] \tag{39}$$

$$= l\underline{p}_t - l\overline{p}_{t'} - \mathrm{log1p}(\exp(l\underline{p}_t - l\overline{p}_{t'}))\,, \tag{40}$$

where $\mathrm{log1p}(\cdot) = \log(\cdot + 1)$.

Following the same argument, we compute $\log \overline{P}(t \mid \mathbf{x})$. Notice that, as the target variable is binary, we have $\underline{P}(t' \mid \mathbf{x}) = 1 - \overline{P}(t \mid \mathbf{x})$ and $\overline{P}(t' \mid \mathbf{x}) = 1 - \underline{P}(t \mid \mathbf{x})$.

The MAP estimation for the credal NB is then defined as:

$$t^* = \operatorname*{argmax}_{t \in \{0,\,1\}} \ \log \underline{P}(t \mid \mathbf{x})\,. \tag{41}$$

If $\underline{P}(t^* \mid \mathbf{x}) > 0.5$, then the classification is said to be *certain*, and we obtain both interval and credal dominance. Otherwise, the classification is considered *uncertain*.