Bayesian Network Structural Consensus via Greedy Min-Cut Analysis

Pablo Torrijos[1,2*], **José M. Puerta**[1,2], **Juan A. Aledo**[1,3], **José A. Gámez**[1,2]

[1]Instituto de Investigación en Informática de Albacete (I3A), Universidad de Castilla-La Mancha, Albacete, Spain
[2]Departamento de Sistemas Informáticos, Universidad de Castilla-La Mancha, Albacete, Spain
[3]Departamento de Matemáticas, Universidad de Castilla-La Mancha, Albacete, Spain

Abstract

This paper presents the Min-Cut Bayesian Network Consensus (MCBNC) algorithm, a greedy method for structural consensus of Bayesian Networks (BNs), with applications in federated learning and model aggregation. MCBNC prunes weak edges from an initial unrestricted fusion using a structural score based on min-cut analysis, integrated into a modified Backward Equivalence Search (BES) phase of the Greedy Equivalence Search (GES) algorithm. The score quantifies edge support across input networks and is computed using max-flow. Unlike methods with fixed treewidth bounds, MCBNC introduces a pruning threshold θ that can be selected post hoc using only structural information. Experiments on real-world BNs show that MCBNC yields sparser, more accurate consensus structures than both canonical fusion and the input networks. The method is scalable, data-agnostic, and well-suited for distributed or federated structural learning of BNs or causal discovery.

Code — https://github.com/ptorrijos99/BayesFL
Datasets — https://doi.org/10.5281/zenodo.14917796
Appendix — https://arxiv.org/abs/2504.00467

Introduction

Bayesian Networks (BNs) (Jensen and Nielsen 2007; Koller and Friedman 2009) are a formalism for modeling uncertainty probabilistically, with widespread applications in domains such as medical diagnosis (McLachlan et al. 2020), bioinformatics (Angelopoulos et al. 2022; Bernaola et al. 2023), and environmental risk assessment (Dai et al. 2024). Their semantic clarity, stemming from the encoding of conditional independencies via Directed Acyclic Graphs (DAGs), makes them particularly attractive for interpretable decision-making (Meekes, Renooij, and van der Gaag 2015). In many scenarios, it is necessary to aggregate multiple BNs, whether elicited from different experts or learned from disjoint datasets, into a single consensus structure. This task, known as *structural fusion* (Peña 2011), aims to consolidate shared independencies while minimizing model redundancy. Both BN learning and fusion are NP-hard (Jensen and Nielsen 2007), and naïve aggregation strategies often lead to complex models with poor inference performance.

A common approach is to compute the union of the input DAGs under a fixed node ordering (Puerta et al. 2021), producing a dense structure that contains all independences supported by at least one input BN. Although this guarantees the definition of structural fusion, it tends to inflate the treewidth (tw) of the resulting network, severely limiting its practical use. The time complexity of exact inference in a BN is exponential in this tw, specifically $O(n \cdot k^{tw+1})$ (Chandrasekaran, Srebro, and Harsha 2008), where n is the number of variables and k the number of states per variable.

To address this, pruning-based methods have been studied. Genetic algorithms have been used to enforce treewidth constraints via edge deletion (Torrijos, Gámez, and Puerta 2024), and more recently, to directly optimize consensus structure quality under user-defined objectives (Torrijos et al. 2025). However, these methods remain computationally expensive and require setting parameters such as target treewidth or stopping criteria, which are difficult to determine without access to data, limiting their application to scenarios such as federated learning (McMahan et al. 2017).

Greedy algorithms offer a scalable alternative, but with clear limitations. Torrijos, Gámez, and Puerta (2024) also proposed a greedy pruning that approximates the unrestricted fusion; Torrijos et al. (2025) used a similar heuristic to mimic the input graphs. Both rely on edge frequency and ignore other structural properties (Koller and Friedman 2009), so they serve only as initializers for genetic algorithms and fail to operate standalone. They also require a fixed treewidth bound: a value set too low removes essential edges, while a high value leaves inference intractable.

We propose a scalable, parameter-light strategy for recovering a consensus structure from input graphs without access to data. Our method begins from the unrestricted fusion obtained using the heuristic node ordering of Puerta et al. (2021) and iteratively prunes edges based on a flow-based structural score that captures edge support across the input networks. This process prunes spurious dependencies without constraining treewidth. The only free parameter is a pruning threshold θ, which can be near-optimally selected a posteriori using only the input graph structures.

Federated learning (McMahan et al. 2017) lets clients train models collaboratively without sharing private data. In

*Corresponding author. Email: Pablo.Torrijos@uclm.es.

the context of BNs, one natural approach (Torrijos, Gámez, and Puerta 2025) is for each client to learn a local structure from its own dataset, which is then aggregated into a global consensus BN. Structural fusion is therefore the critical step, performed without data or a gold standard. Experiments in this setting confirm that our method, *Min-Cut Bayesian Network Consensus* (MCBNC), consistently produces consensus structures that are not only sparser and more interpretable than those from canonical fusion but also more faithful to the underlying dependency structure than the input networks themselves on average.

Contributions. The principal contributions are:

- A max-flow–based score to quantify edge support.
- Integration of this score into the Backward Equivalence Search (BES) phase of the Greedy Equivalence Search (GES) algorithm to prune edges within the Markov equivalence class of the fused network.
- An adaptive pruning rule with a single threshold θ, selected post hoc using only input graphs.

Paper organization. The paper proceeds as follows. **Background** reviews key concepts in BN fusion and flow-based analysis. **Proposal** introduces the MCBNC algorithm and its theoretical foundations. **Experimental Methodology** details the evaluation setup. **Experimental Results** report results on real and synthetic networks. **Conclusions** summarize findings and future directions.

Preliminaries

Bayesian Networks

A Bayesian Network (BN) is a pair $B = (G, P)$, where $G = (V, E)$ is a directed acyclic graph (DAG) representing conditional (in)dependences over variables $V = \{v_1, \ldots, v_n\}$, and P is a set of probability distributions that factorizes as

$$\mathbb{P}(V) = \prod_{i=1}^{n} \mathbb{P}\big(v_i \mid \mathbf{Pa}_G(v_i)\big), \tag{1}$$

where $\mathbf{Pa}_G(v_i)$ denotes the parent set of v_i in G. The graph G encodes conditional independencies $I(G)$ via *d-separation* (Koller and Friedman 2009). A DAG G is an I-map of G' when $I(G) \subseteq I(G')$ and is *minimal* if removing any arc destroys this property. DAGs that encode the same $I(G)$ form a Markov equivalence class, representable by a Completed Partially Directed Acyclic Graph (CPDAG) $\mathcal{G}$ (Chickering 2002). In $\mathcal{G}$, directed edges appear when their orientation is invariant across all equivalent DAGs; undirected edges denote ambiguity.

Treewidth. Let $\widetilde{G}$ be the moral graph of G (all parents of each node joined and edges made undirected). The treewidth $\mathrm{tw}(G)$ is the size of the largest clique[1] in an optimal triangulation of $\widetilde{G}$ minus one. Exact inference is $O\big(n\, k^{\mathrm{tw}(G)+1}\big)$, where k is the maximum state count per variable (Chandrasekaran, Srebro, and Harsha 2008); low treewidth is therefore essential for BN tractability and usability.

[1] A clique is a fully connected node subset.

Structural Fusion of Bayesian Networks

Let $\{G_i = (V, E_i)\}_{i=1}^{r}$ be DAGs over a shared variable set V. A common structural fusion strategy (Peña 2011; Puerta et al. 2021) applies a total node ordering σ to each G_i, producing acyclic DAGs $\{G_i^{\sigma}\}_{i=1}^{r}$ where all parents of a node precede it. The fused DAG is then

$$G^{+} = (V, E^{+}), \qquad E^{+} = \bigcup_{i=1}^{r} E_i^{\sigma}. \tag{2}$$

This union is guaranteed to be acyclic and is a minimal I-map of the intersection $\bigcap_i I(G_i^{\sigma})$. The final density of G^{+} depends strongly on the ordering σ, since some orderings induce fewer edges when reorienting the G_i. Finding the optimal σ is NP-hard, so we adopt the heuristic from (Puerta et al. 2021), which gives near-optimal orderings in practice.

From fusion to consensus. Strict fusion retains all dependencies present in any input, often producing dense graphs with high treewidth, especially when the G_i are heterogeneous. To address this, we define a *consensus* DAG $G^{*} = (V, E^{*})$ that maximizes a structural score:

$$E^{*} = \arg\max_{E' \in \mathcal{E}} \sum_{e \in E'} \psi(e), \tag{3}$$

where $\mathcal{E}$ is a search space (e.g., subsets of E^{+} or possible edges on V), and $\psi(e)$ quantifies how strongly edge e is supported across the input networks. This idea was formalized in Torrijos et al. (2025) as an alternative to canonical fusion, enabling more interpretable and tractable structures.

Backward Equivalence Search (BES)

Greedy Equivalence Search (GES) is a two-phase algorithm for BN structure learning (Chickering 2002). It first adds edges in a forward phase and then removes them in a backward phase, Backward Equivalence Search (BES). Both phases operate over Markov-equivalent classes and use a decomposable score, such as Bayesian Dirichlet equivalent uniform (BDeu), to guide edge modifications. BES iteratively deletes the edge that gives the most significant score improvement. Formally, given a DAG $G = (V, E)$, data D and the score $f(G : D)$, BES replaces G by

$$G' = \arg\max_{e \in E} f\big(G \setminus \{e\} : D\big), \tag{4}$$

and stops when no deletion increases the score. Its DELETE operator (Chickering 2002) will be reused by our method.

Min-cut and max-flow. Let $D = (V, E)$ be a directed graph with non-negative capacities $c : E \to \mathbb{R}^{+}$. For a source s and sink t, a cut (S, T) satisfies $s \in S$, $t \in T$, $S \cup T = V$, $S \cap T = \emptyset$, and has capacity

$$\mathrm{cap}(S, T) = \sum_{u \in S,\, v \in T} c(u \to v). \tag{5}$$

The *min-cut* problem seeks the cut of minimum capacity. The *max-flow* problem finds a flow $f : E \to \mathbb{R}^{+}$ that respects capacities and flow conservation and maximises

$$\mathrm{val}(f) = \sum_{e \in \delta^{+}(s)} f(e). \tag{6}$$

The Max-Flow Min-Cut Theorem (Ahuja, Magnanti, and Orlin 1993) states

$$\max_f \text{val}(f) = \min_{(S,T)} \text{cap}(S,T). \tag{7}$$

Ford-Fulkerson algorithm. Any polynomial-time max-flow routine can be used. We employ the classical Ford-Fulkerson augmenting-path algorithm (Ford and Fulkerson 1956) for its simplicity. Implementation details are standard; refer to the Technical Appendix (Sec. D) for details.

Method: Min-Cut Bayesian Network Consensus (MCBNC)

Structural fusion methods (e.g., (Puerta et al. 2021)) compute a fused DAG G^+ that retains all (in)dependencies in the input BNs $\{B_i\}_{i=1}^r$ with structures $\{G_i = (V, E_i)\}_{i=1}^r$. While correct by construction, G^+ is often dense and yields high treewidth, which limits its usability. Our method, *Min-Cut Bayesian Network Consensus* (MCBNC), addresses this by iteratively pruning weakly supported edges from G^+. The approach builds on the Backward Equivalence Search (BES) phase of Greedy Equivalence Search (GES) (Chickering 2002), replacing its likelihood-based scoring with a structural score based on the max-flow min-cut algorithm. This score quantifies the support of each edge across the input graphs and enables parameterized pruning using a threshold θ. The intuition is that an edge $u \to v$ is critical only if its removal would disconnect u and v in the moralized ancestral subgraphs of many input DAGs. If many alternative paths exist, the min-cut is large, indicating the edge is redundant. Pruning such weakly supported edges simplifies fusion while preserving consensus dependencies.

Before pruning, G^+ is converted to its CPDAG $\mathcal{G}^+$ to ensure compatibility with BES operators such as DELETE (Chickering 2002). The complete procedure is summarized in Alg. 1, with each component detailed in the subsections below. A simple example of the algorithm's execution is provided in the Technical Appendix (Sec. E).

Algorithm 1 Min-Cut Bayesian Network Consensus

Require: Input DAGs $\{G_i = (V, E_i)\}_{i=1}^r$, threshold θ, maximum subset size $k_{\max}$
Ensure: Consensus DAG G^*

$\sigma \leftarrow$ ORDERING($\{G_i\}$) ▷ (Puerta et al. 2021)
for $i = 1$ to r **do**
  $G_i^\sigma \leftarrow$ MINIMALIMAP(G_i, σ) ▷ (Peña 2011)
end for
$G^+ \leftarrow (V, \bigcup_i E_i^\sigma)$ ▷ Unrestricted fusion
$\mathcal{G} \leftarrow$ DAGTOCPDAG(G^+) ▷ (Chickering 2002)
while true do
  $(e^*, H^*, \Psi^*, \mathcal{C}^*) \leftarrow$ BESTEDGE($\mathcal{G}, \{G_i\}_{i=1}^r, k_{\max}$)
  if $\Psi^* > \theta$ **then break**
  end if
  $\mathcal{G} \leftarrow$ DELETE($\mathcal{G}, e^*, H^*$) ▷ (Chickering 2002)
  $\{G_i \leftarrow G_i \setminus \mathcal{C}_i\}_{i=1}^r$ ▷ Remove cut edges
end while
$G^* \leftarrow$ PDAGTODAG($\mathcal{G}$) ▷ A DAG consistent with $\mathcal{G}$
return G^*

Edge Criticality via Min-Cut

MCBNC prioritizes edge removals that preserve key dependencies while reducing graph complexity. To guide this, a *criticality score* $\Psi^H_{(u\to v)}$ is computed from flow separation in the moralized input DAGs. The score quantifies the structural relevance of each edge $e = (u \to v)$ in the fused CPDAG $\mathcal{G}^+$. Following Chickering (2002), deletions must preserve the Markov equivalence class. For each edge $e = (u \to v)$ in $\mathcal{G}^+$, the set of valid conditioning nodes is:

$$\mathcal{N}_{uv} = \{w \mid w \to v \text{ in } \mathcal{G}^+ \text{ and } w - u \text{ is undirected in } \mathcal{G}^+\}. \tag{8}$$

Given a candidate subset $H \subseteq \mathcal{N}_{uv}$, the criticality score $\Psi^H_{(u\to v)}$ is computed as follows (Alg. 2):

1. For each input DAG $\{G_i\}_{i=1}^r$, extract the ancestral subgraph[2] of $\{u, v\} \cup H$, moralize it, and remove all nodes in H, yielding the conditioned graphs $\{\widetilde{G}_i^H\}_{i=1}^r$.
2. On each conditioned graph $\{\widetilde{G}_i^H\}_{i=1}^r$, compute the size of the minimum cut separating u and v using the Ford-Fulkerson algorithm (Ford and Fulkerson 1956).
3. Return the average cut size across all graphs, which defines the criticality score $\Psi^H_{(u\to v)}$.

Algorithm 2 CRITICALITY

function CRITICALITY($(u \to v)$, $\{G_i\}_{i=1}^r$, H)
  for $i = 1$ **to** r **do**
    $A_i \leftarrow$ ANCESTRALSUBGRAPH$\big(G_i, \{u, v\} \cup H\big)$
    $\widetilde{G}_i^H \leftarrow$ MORALIZE$(A_i) \setminus H$
    $S_i^H \leftarrow$ MINCUT$\big(\widetilde{G}_i^H, u, v\big)$
  end for
  $\Psi^H_{(u\to v)} \leftarrow \frac{1}{r}\sum_{i=1}^r |S_i^H|$
  $\mathcal{C}^H_{(u\to v)} \leftarrow \bigcup_{i=1}^r S_i^H$
  return $\big(\Psi^H_{(u\to v)}, \mathcal{C}^H_{(u\to v)}\big)$
end function

Edges with lower $\Psi^H_{(u\to v)}$ contribute less to the structural integrity of the fused network and are prioritized for removal. This score-based strategy replaces likelihood-based criteria and avoids fixed structural constraints.

Greedy Edge Search

MCBNC performs edge deletion through a greedy search over the space of Markov equivalence classes, following the Backward Equivalence Search (BES) strategy from GES (Chickering 2002). Pruning operates on the CPDAG $\mathcal{G}^+$, where edges can be directed or undirected. Undirected edges are evaluated in both orientations $(u \to v)$ and $(v \to u)$, ensuring that all valid deletion candidates are considered.

The function BESTEDGE (Alg. 3) selects, at each iteration, the least critical edge based on its structural support. For each arc $(u \to v)$, the procedure is as follows:

[2]The ancestral subgraph of a set S in a DAG G is the subgraph induced by all nodes from which there exists a directed path to some node in S, including the nodes in S themselves.

1. Identify the valid conditioning set $\mathcal{N}_{uv}$ of nodes that are parents of v and share an undirected edge with u in $\mathcal{G}^+$.
2. Generate all subsets $H \subseteq \mathcal{N}_{uv}$ of size at most $k_{\max}$, where $k_{\max}$ is a user-defined pruning budget.
3. For each H, compute the criticality score $\Psi^H_{(u\to v)}$ using the method on Alg. 2.
4. Select the pair (e^*, H^*) minimizing the score and return the edge $e^* = (u \to v)$, its score $\Psi^{H^*}_{e^*}$, the conditioning set H^*, and the union of cut sets $\mathcal{C}^{H^*}_{e^*}$.

Algorithm 3 BESTEDGE

function BESTEDGE($\mathcal{G}, \{G_i\}_{i=1}^r, k_{\max}$)
  $\Psi^* \leftarrow \infty$
  for all $(u \to v) \in \mathcal{G}$ **do** ▷ $u{-}v \Rightarrow u{\to}v, v{\to}u$
    $\mathcal{N}_{uv} \leftarrow \{w \mid w \to v$ and $w{-}u$ undirected in $\mathcal{G}\}$
    for all $H \subseteq \mathcal{N}_{uv},\ |H| \le k_{\max}$ **do**
      $S \leftarrow (\mathcal{N}_{uv} \setminus H) \cup (\text{PARENTS}(v, \mathcal{G}) \setminus \{u\})$
      $(\Psi, \mathcal{C}) \leftarrow \text{CRITICALITY}((u{\to}v), \{G_i\}, S)$
      if $\Psi < \Psi^*$ **then**
        $(e^*, H^*, \Psi^*, \mathcal{C}^*) \leftarrow ((u{\to}v), H, \Psi, \mathcal{C})$
      end if
    end for
  end for
  return $(e^*, H^*, \Psi^*, \mathcal{C}^*)$
end function

Main Iterative Pruning Scheme

MCBNC removes edges from $\mathcal{G}^+$ greedily, following the BES strategy from GES (Chickering 2002). At each step, it deletes the edge with the lowest criticality score $\Psi^H_{(u\to v)}$, provided $\Psi^H_{(u\to v)} \le \theta$. The process stops when no such edge remains. Alternatively, θ, the algorithm can run until $\mathcal{G}^+$ is empty, retaining the structure with minimal average structural distance to the inputs. This enables parameter-free model selection, avoiding the need for predefined treewidth bounds. The complete procedure is summarized in Alg. 1:

1. Fuse the input DAGs into G^+ using a heuristic ordering as in Puerta et al. (2021).
2. Convert G^+ into its CPDAG $\mathcal{G}^+$ to operate within the equivalence class using Chickering (2002).
3. Repeatedly:
   (a) Use BESTEDGE (Alg. 3) to find the edge e^* and conditioning set H^* minimizing $\Psi^H_{e^*}$.
   (b) If $\Psi^H_{e^*} > \theta$, stop.
   (c) Remove e^* using DELETE (Chickering 2002), update the graphs, and convert to CPDAG.

Implementation assumptions. All edge capacities are assumed to be one. For each candidate edge, all conditioning subsets $H \subseteq \mathcal{N}_{uv}$ of size at most $k_{\max}$ are enumerated. This is feasible since $|\mathcal{N}_{uv}|$ is typically small, and $k_{\max}$ is fixed. The choice of max-flow algorithm is flexible; any correct implementation (e.g., Edmonds-Karp, Dinic) can be used, as the score depends only on the size of the minimum cut. Acyclicity is preserved by applying the DELETE operator within the Markov equivalence class.

Properties

This section states key properties of MCBNC.

Lemma 1 (Monotonicity of the criticality score)**.** *Let $\Psi_e^{(t)}$ be the criticality score of edge e after the t-th deletion. Then $\Psi_e^{(t+1)} \ge \Psi_e^{(t)}$ for every remaining edge e.*

Proof. Deleting an edge can only remove paths in the ancestral moral graphs used for computing criticality. Since the min-cut size is determined by the number of edge-disjoint paths between u and v, its value cannot increase. Hence, the score is monotonic and non-increasing. □

Corollary 2 (Score interpretation)**.** *Let $e = (u \to v)$ appear in exactly k of the r input DAGs and suppose all u-v paths in those DAGs include e. Then $\Psi_e = k/r$ and:*

$$\theta < k/r \Rightarrow e \text{ is retained}, \quad \theta \ge k/r \Rightarrow e \text{ is removed}.$$

Lemma 3 (Complexity of MCBNC with Ford-Fulkerson)**.** *Let r be the number of input DAGs, $m = |E^+_\sigma|$ the number of edges in the unrestricted fusion, and $k_{\max}$ the conditioning-set cap. With unit capacities and Ford-Fulkerson for min-cut, MCBNC runs in $O\big(r\, m^3\, 2^{k_{\max}}\big)$ time and $O(r\, m)$ space.*

Proof. Each min-cut takes $O(m^2)$ time. A criticality score requires r min-cuts, costing $O(r\, m^2)$. For $2^{k_{\max}}$ subsets per edge and m edges per iteration, BESTEDGE costs $O(r\, m^2\, 2^{k_{\max}})$. The greedy loop runs at most m iterations, giving total time $O(r\, m^3\, 2^{k_{\max}})$. Memory is dominated by the CPDAG and r DAGs, each with $O(m)$ edges. □

Experimental Methodology

We evaluate MCBNC in both synthetic and realistic fusion settings. In both cases, the goal is to recover a consensus DAG G^* that approximates a known gold-standard Bayesian network G_{gs}. Let $\{G_i\}_{i=1}^r$ denote the input DAGs, obtained either by structural perturbation of G_{gs} (synthetic setup) or by learning from data sampled from G_{gs} (federated setup).

As a sanity check and to replicate prior work, we first follow and extend the synthetic setup of (Puerta et al. 2021), where each G_i is derived by randomly perturbing G_{gs}. In this idealized case, MCBNC consistently reconstructs G_{gs} with near-zero Structural Moral Hamming Distance (SMHD), even for large networks. These results confirm correctness and are reported in the Technical Appendix (Sec. B).

We then evaluate MCBNC in a more realistic and challenging federated setting. Each of the $r \in \{5, 10, 20, 30, 50, 100\}$ clients receive a private dataset D_i of 5000 independent and identically distributed (i.i.d.) samples from G_{gs} and learns a local DAG G_i using the GES algorithm. The fusion operates solely on the structures $\{G_i\}_{i=1}^r$ without accessing the underlying data $\{D_i\}_{i=1}^r$. The goal is for the consensus network G^* to recover the dependency structure of G_{gs}, despite the variability introduced by limited-data learning.

As gold standards, we utilize 15 benchmark networks from the BNLEARN repository (Scutari 2010), which cover a broad range of sizes and topologies (see Table 1).

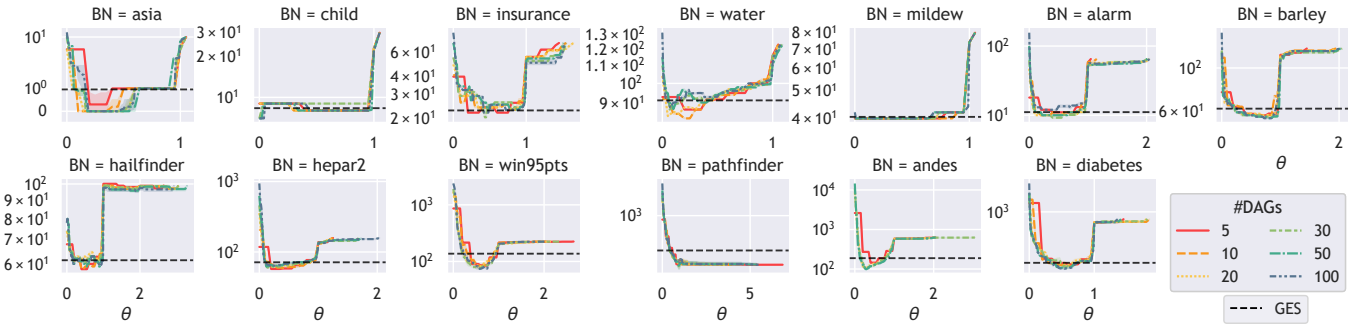

Figure 1: Mean SMHD to the gold-standard BN G_{gs} across thresholds θ for each BN. Leftmost point: full fusion G^{+}. Rightmost: empty DAG $\emptyset$. Horizontal line: average SMHD of input BNs from GES to G_{gs}. Lower is better.

NETWORK	$\lvert V\rvert$	$\lvert E\rvert$	NETWORK	$\lvert V\rvert$	$\lvert E\rvert$	NETWORK	$\lvert V\rvert$	$\lvert E\rvert$
ASIA	8	8	MILDEW	35	46	WIN95PTS	76	112
SACHS	11	17	ALARM	37	46	PATHFINDER	109	195
CHILD	20	25	BARLEY	48	84	ANDES	223	338
INSURANCE	27	52	HAILFINDER	56	66	DIABETES	413	602
WATER	32	66	HEPAR2	70	123	PIGS	441	592

Table 1: Benchmark Bayesian networks (nodes/edges).

Experimental Protocol. For each benchmark network[3] and each $r \in \{5, 10, 20, 30, 50, 100\}$:

(1) A collection of r datasets $\{D_i\}_{i=1}^{r}$ is generated by drawing 5000 i.i.d. samples from the gold-standard BN. Each D_i is used to learn a local DAG G_i via GES.

(2) The input structures $\{G_i\}_{i=1}^{r}$ are fused into a DAG G^{+} using the fusion method of Puerta et al. (2021).

(3) MCBNC is executed from G^{+}, iteratively pruning edges. The algorithm produces the full trajectory $\{G^{*}(\theta)\}$ for all thresholds θ in a single run.

(4) Steps (2)–(3) are repeated 10 times per configuration, using the same input DAGs, to assess robustness to algorithmic randomness (e.g., tie-breaking, ordering).

(5) Each consensus DAG $G^{*}(\theta)$ is evaluated using multiple structural and data-based metrics.

Conditioning set size. We fix the conditioning-set cap to $k_{\max} = 10$ as an internal constant; it is not a user-tuned parameter. In practice, conditioning sets are small because they derive from nodes adjacent to both endpoints of an undirected edge in the current CPDAG, and their size shrinks as pruning progresses. Ablation results in Technical Appendix (Sec. C.4) confirm that varying $k_{\max}$ has negligible impact on consensus quality or runtime, as large sets are rarely generated.

Evaluation Metrics. Each consensus DAG $G^{*}(\theta)$ is assessed using the following criteria:

- **SMHD:** The Structural Moral Hamming Distance (Kim and Kim 2019; Torrijos, Gámez, and Puerta 2024) quantifies structural differences after moralization. We compute the mean SMHD to the gold-standard BN (measuring fidelity) and to the input DAGs (measuring consensus). Lower values are better.
- **BDeu Score:** The Bayesian Dirichlet equivalent uniform score (Chickering 2002) quantifies data likelihood given the structure. MCBNC ignores this criterion during pruning; we report it only for reference (larger is better).
- **Treewidth:** Indicates structural complexity and governs the cost of exact inference. Lower treewidths are desirable because they imply more tractable models.

Technical Appendix (Sec. A) provides extended metric definitions and additional structural indicators.

Implementation and Reproducibility

All code was implemented in Java (OpenJDK 17) using the TETRAD 7.6.5 causal inference library.[4] Structure learning was performed with GES. All real-world networks were obtained from the BNLEARN repository (see Table 1). Experiments were run on Intel Xeon E5-2650 (8 cores) with 32 GB RAM per run. To ensure full reproducibility, we provide all source code, experiment scripts, and preprocessed datasets on GitHub.[5] The datasets are also archived on Zenodo.[6] Statistical tests were carried out using the EXREPORT package (Arias and Cozar 2016) for R.

Experimental Results

We present the results of applying MCBNC in the federated learning scenario. Each figure plots performance metrics as a function of the fusion threshold θ. The leftmost point corresponds to the initial fusion G^{+} (Puerta et al. 2021), while the rightmost reflects the empty network.

Structural Accuracy (SMHD)

Fig. 1 shows how SMHD of G^{*} to the gold-standard BN G_{gs} varies with the pruning threshold θ (from G^{+} on $\theta = 0$ to $\emptyset$ on the last θ). In almost all cases, G^{+} yields worse SMHD

[3] BNs SACHS and PIGS are omitted from the main plots because GES already yields their gold-standard DAGs. Consequently, the fusion G^{+} is optimal, and MCBNC deletes no edges for $\theta < 1$. Detailed results appear in Technical Appendix (Sec. C.3).

[4] https://github.com/cmu-phil/tetrad/releases/tag/v7.6.5

[5] https://github.com/ptorrijos99/BayesFL

[6] https://doi.org/10.5281/zenodo.14917796

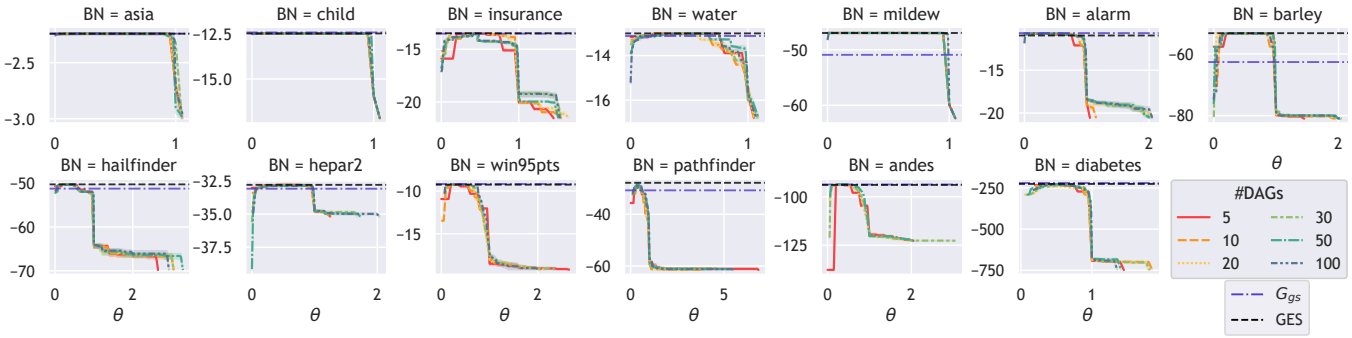

Figure 2: Mean BDeu score across thresholds θ for each BN. Leftmost point: full fusion G^+. Rightmost: empty DAG $\emptyset$. Horizontal lines: average of input BNs from GES (black) and gold-standard BN (purple). Higher is better.

than even the empty DAG, confirming that unrestricted fusion accumulates spurious dependencies and the need for consensus fusions. Applying MCBNC yields steep SMHD reductions[7], particularly in large networks like ANDES or DIABETES, where improvements over G^+ span up to two orders of magnitude. Gains relative to the GES-generated input DAGs are also notable, as MCBNC removes dataset-specific artifacts and consolidates shared dependencies, resulting in BNs that are more similar to G_{gs}. Performance remains stable across a broad range of θ values, with over-pruning (and SMHD degradation) occurring near $\theta = 1$.

Data Fit (BDeu Score)

Fig. 2 reports the BDeu scores of the consensus networks across different values of θ. GES optimizes BDeu directly, so its input DAGs perform strongly. MCBNC, by contrast, neither accesses the data nor optimizes any likelihood-based objective. Still, it achieves scores comparable to (and occasionally exceeding) those of the input networks. In some cases, such as the BARLEY and MILDEW BNs, even the gold-standard structure yields lower BDeu. This well-known phenomenon arises because sparser graphs, which correctly reflect the true dependencies, may underfit finite datasets. In PATHFINDER, MCBNC again outperforms the gold standard in BDeu, but this does not imply a better structure: SMHD remains high (Fig. 1), confirming that BDeu and structural accuracy do not always align. Overall, MCBNC achieves competitive BDeu scores despite being data-agnostic. Still, selecting an appropriate fusion threshold θ is crucial.

Choosing the Fusion Threshold θ

Detailed SMHD-BDeu curves for each client count $r \in \{5, 10, 20, 30, 50, 100\}$ are reported in Technical Appendix (Sec. C.2). These plots show that the threshold θ minimizing SMHD to the input DAGs also tends to maximize structural agreement with the gold-standard network and yields strong BDeu scores. This supports a practical selection strategy:

[7]An exception is the PATHFINDER BN, where SMHD improves monotonically even as the network is pruned to near emptiness. This reflects a structural mismatch in the input DAGs, as GES fails to recover the underlying semi-Naive Bayes structure. This limitation is known in the literature (Laborda et al. 2024).

METRIC	METHOD	RANK	p-VALUE	W / T / L
SMHD	**MCBNC** (G^*)	**1.40**	—	—
	GES $(\{G_i\}_{i=1}^r)$	1.91	6.95×10^{-4}	61 / 13 / 16
	Fusion (G^+)	2.69	7.68×10^{-18}	68 / 17 / 5
BDEU	GES $(\{G_i\}_{i=1}^r)$	**1.58**	—	—
	MCBNC (G^*)	1.70	$\mathbf{4.12 \times 10^{-1}}$	48 / 9 / 33
	Fusion (G^+)	2.72	3.25×10^{-14}	71 / 9 / 10

Table 2: Statistical comparison over 15 BNs and six client counts (90 cases). Lower rank is better. p-values refer to Holm's procedure against the top-ranked method; bold values indicate non-rejection of H_0 at $\alpha = 0.01$.

set θ post hoc to minimize the mean SMHD to the input graphs. This criterion requires no access to data or ground truth, making it suitable for realistic scenarios such as federated learning. Rather than displaying the six SMHD–BDeu curves, we summarize the evidence statistically below.

Method. For each benchmark BN and each r, we extracted the consensus DAG $G^*(\theta)$ on the point θ that minimized SMHD to the input GES DAGs $\{G_i\}_{i=1}^r$. Three algorithms were compared: (i) MCBNC (G^*) at the selected θ, (ii) the average of the r GES DAGs, and (iii) the unrestricted fusion G^+. Ranks over benchmarks were analysed with the Friedman test (Friedman 1940) to assess whether all methods perform equally. If the null hypothesis was rejected, pairwise differences were tested using Holm's post-hoc correction (Holm 1979). Both tests used $\alpha = 0.01$, following standard practice (Demsar 2006; García and Herrera 2008).

Interpretation. The Friedman test rejects the null hypothesis of equal methods for both metrics: $p = 2.32 \times 10^{-17}$ for SMHD and $p = 3.66 \times 10^{-16}$ for BDeu. Holm's post-hoc analysis (Table 2) confirms that, for SMHD, MCBNC significantly outperforms both the GES average and the unrestricted fusion. Among the ties, 12 correspond to SACHS and PIGS, where GES already recovers G_{gs} and no structural improvement is possible. The rest occur in small networks, where differences are minor. For BDeu, MCBNC and GES are statistically indistinguishable ($p \approx 0.41$), while both significantly outperform the unrestricted fusion. This is ex-

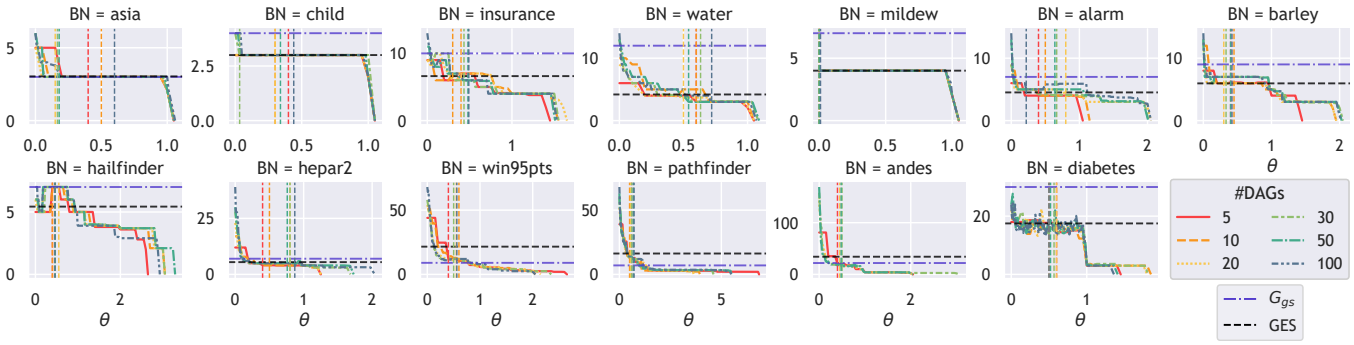

Figure 3: Mean treewidth across pruning thresholds θ for each BN. Dashed lines: selected θ for each #DAGs based on SMHD w.r.t. input BNs. Horizontal lines: average of input BNs from GES (black) and gold-standard BN (purple).

pected: GES optimizes and overfits BDeu, whereas MCBNC still yields competitive likelihood. These results confirm that selecting θ by minimizing SMHD to the input GES DAGs yields consensus networks that are structurally faithful and competitive in terms of data fit.

Structural Properties of the Fused Networks

Fig. 3 plots the treewidth of the consensus BNs as θ varies (edge-count curves are in Technical Appendix C.1). Pruning with small θ eliminates many weak edges, producing an immediate and drastic drop in treewidth. For $\theta \in [0.2, 0.8]$ the curve flattens: MCBNC has removed most surplus edges yet still preserves the backbone of dependencies. Beyond $\theta \approx 0.9$, relevant edges vanish and treewidth falls again, mirroring the rise in SMHD. The vertical dotted lines mark the selected θ for each number of clients. At those points, the consensus graphs are never denser (and are frequently sparser) than both the gold-standard and the individual GES models, despite matching or surpassing them in SMHD. Networks such as WIN95PTS illustrate the benefit: the treewidth drops from approximately 20 to around 10, while the mean SMHD to the gold standard improves by 58.7% (Fig. 1).

Runtime Comparison with Prior Methods

Figure 4 shows the runtime of MCBNC compared to the genetic fusion algorithms from Torrijos, Gámez, and Puerta (2024); Torrijos et al. (2025), using the same networks and number of input DAGs as in those studies. The algorithm in Torrijos, Gámez, and Puerta (2024) searches over the set E_{G^+}, corresponding to arcs in the unrestricted fusion. The method in Torrijos et al. (2025) generalizes this by operating over $E_{\mathbb{G}}$ (all input edges, with repetition) or $E^*_{\mathbb{G}}$ (without repetition), depending on the chromosome encoding. Despite these differences, all genetic variants show similar scaling. MCBNC is several orders of magnitude faster, making it impractical to replicate our complete evaluation with these algorithms. The reliance on a fixed treewidth in the other methods complicates fair comparisons, as no unique treewidth target applies across networks or aggregation levels. Complete runtime results for MCBNC are provided in the Technical Appendix (Sec. C.1).

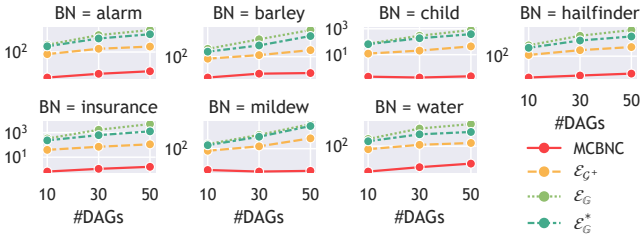

Figure 4: Total execution time (s) vs. number of input DAGs.

Conclusions

This work introduced the Min-Cut Bayesian Network Consensus (MCBNC) algorithm for structure-level fusion of Bayesian networks. MCBNC overcomes limitations of existing fusion methods by pruning non-essential edges through a backward search guided by min-cut analysis. Unlike unrestricted fusion (Puerta et al. 2021), which preserves all independencies at the cost of excessive complexity, or treewidth-bounded approaches that require a user-defined structural constraint (Torrijos, Gámez, and Puerta 2024; Torrijos et al. 2025), MCBNC offers an interpretable and tunable alternative controlled by a single threshold θ. Empirically, MCBNC consistently yields consensus networks that are not only simpler and more tractable, but also structurally more faithful to the underlying dependency model, as measured by SMHD. Remarkably, it achieves this without accessing any data, relying solely on the input graph structures. Moreover, the pruning threshold θ can be selected near-optimally in a post-hoc manner using only structural information (e.g., by minimizing SMHD to the input DAGs), making MCBNC particularly well-suited to realistic, data-scarce settings such as federated learning and privacy-preserving causal discovery.

Immediate future work includes extending the flow-based score to mixed or evolving variable sets, studying robustness under non-i.i.d. client distributions, and comparing against frequency-based consensus baselines. Embedding MCBNC into advanced federated frameworks, such as FedGES (Torrijos, Gámez, and Puerta 2025), could further enhance its utility in applications ranging from distributed diagnostics to causal inference under privacy constraints.

Acknowledgements

This work was supported by SBPLY/21/180225/000062 (Junta de Comunidades de Castilla-La Mancha and ERDF A way of making Europe); PID2022-139293NB-C32 (MICIU/AEI/10.13039/501100011033 and ERDF, EU); FPU21/01074 (MICIU/AEI/10.13039/501100011033 and ESF+); 2025-GRIN-38476 (Universidad de Castilla-La Mancha and ERDF A way of making Europe); TED2021-131291B-I00 (MICIU/AEI/10.13039/501100011033 and European Union NextGenerationEU/PRTR).

References

Ahuja, R. K.; Magnanti, T. L.; and Orlin, J. B. 1993. *Network flows: theory, algorithms, and applications*. USA: Prentice-Hall, Inc.

Angelopoulos, N.; Chatzipli, A.; Nangalia, J.; Maura, F.; and Campbell, P. J. 2022. Bayesian networks elucidate complex genomic landscapes in cancer. *Communications Biology*, 5(1).

Arias, J.; and Cozar, J. 2016. *exreport: Fast, Reliable and Elegant Reproducible Research*. R package version 0.4.1.

Bernaola, N.; Michiels, M.; Larrañaga, P.; and Bielza, C. 2023. Learning massive interpretable gene regulatory networks of the human brain by merging Bayesian networks. *PLOS Computational Biology*, 19(12): e1011443.

Chandrasekaran, V.; Srebro, N.; and Harsha, P. 2008. Complexity of inference in graphical models. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI'08, 70–78. Arlington, Virginia, USA: AUAI Press.

Chickering, D. M. 2002. Optimal Structure Identification With Greedy Search. *Journal of Machine Learning Research*, 3(Nov): 507–554.

Dai, H.; Ju, J.; Gui, D.; Zhu, Y.; Ye, M.; Liu, Y.; Cui, J.; and Hu, B. X. 2024. A two-step Bayesian network-based process sensitivity analysi s for complex nitrogen reactive transport modeling. *Journal of Hydrology*, 632: 130903.

Demsar, J. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7: 1–30.

Ford, L. R.; and Fulkerson, D. R. 1956. Maximal Flow Through a Network. *Canadian Journal of Mathematics*, 8: 399–404.

Friedman, M. 1940. A Comparison of Alternative Tests of Significance for the Problem of m Rankings. *The Annals of Mathematical Statistics*, 11(1): 86–92.

García, S.; and Herrera, F. 2008. An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons. *Journal of Machine Learning Research*, 9: 2677–2694.

Holm, S. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6: 65–70.

Jensen, F. V.; and Nielsen, T. D. 2007. *Bayesian Networks and Decision Graphs*. Springer New York, 2nd edition.

Kim, G.-H.; and Kim, S.-H. 2019. Marginal information for structure learning. *Statistics and Computing*, 30(2): 331–349.

Koller, D.; and Friedman, N. 2009. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press.

Laborda, J. D.; Torrijos, P.; Puerta, J. M.; and Gámez, J. A. 2024. Parallel structural learning of Bayesian networks: Iterative divide and conquer algorithm based on structural fusion. *Knowledge-Based Systems*, 296: 111840.

McLachlan, S.; Dube, K.; Hitman, G. A.; Fenton, N. E.; and Kyrimi, E. 2020. Bayesian networks in healthcare: Distribution by medical condition. *Artificial Intelligence in Medicine*, 107: 101912.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and Arcas, B. A. y. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Singh, A.; and Zhu, J., eds., *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, 1273–1282. PMLR.

Meekes, M.; Renooij, S.; and van der Gaag, L. C. 2015. Relevance of Evidence in Bayesian Networks. In *ECSQARU-2015*, volume 9161 of *Lecture Notes in Computer Science*, 366–375. Springer.

Peña, J. 2011. Finding Consensus Bayesian Network Structures. *The Journal of Artificial Intelligence Research (JAIR)*, 42.

Puerta, J. M.; Aledo, J. A.; Gámez, J. A.; and Laborda, J. D. 2021. Efficient and accurate structural fusion of Bayesian networks. *Information Fusion*, 66: 155–169.

Scutari, M. 2010. Learning Bayesian Networks with the bnlearn Package. *Journal of Statistical Software*, 35(3): 1–22.

Torrijos, P.; Gámez, J. A.; and Puerta, J. M. 2025. FedGES: A Federated Learning Approach for Bayesian Network Structure Learning. In Pedreschi, D.; Monreale, A.; Guidotti, R.; Pellungrini, R.; and Naretto, F., eds., *Discovery Science – DS 2024*. Cham: Springer Nature Switzerland.

Torrijos, P.; Gámez, J. A.; Puerta, J. M.; and Aledo, J. A. 2025. Genetic Algorithms for Tractable Bayesian Network Fusion via Pre-Fusion Edge Pruning. In *Proceedings of the Genetic and Evolutionary Computation Conference 2025*, 481–489. New York, NY, USA: Association for Computing Machinery.

Torrijos, P.; Gámez, J. A.; and Puerta, J. M. 2024. Structural Fusion of Bayesian Networks with Limited Treewidth Using Genetic Algorithms. In *2024 IEEE Congress on Evolutionary Computation (CEC)*, volume 3, 1–8. IEEE.