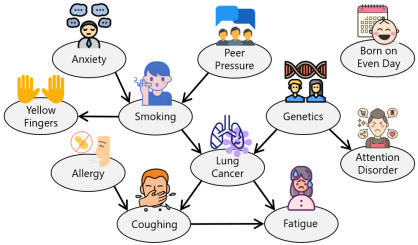
DAG-PnP: Plug-and-play causal discovery with diffusion priors

Nicolas Zilberstein[†], Ali Azizpour[†], Gonzalo Mateos[⋆], Santiago Segarra[†]

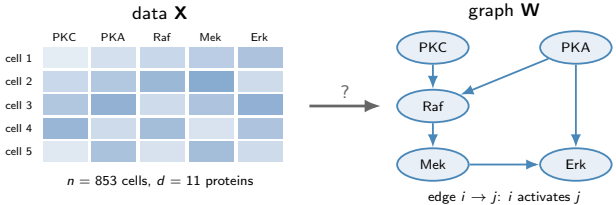
[†] Rice University, USA [⋆] University of Rochester, USA

Asilomar Conference on Signals, Systems, and Computers 2026
Pacific Grove, CA

- Causal discovery: learn *who drives whom* from data
  - $\Rightarrow$ Predict the effect of interventions $\rightarrow$ health, biology, economics



Health: does smoking cause lung cancer?



Biology: protein signaling in single cells [Sachs 2005]

- Typically only **observational** data, and often few samples
  - $\Rightarrow$ The data alone may not pin down the graph $\rightarrow$ need structural priors
  - $\Rightarrow$ We bring a learned prior over DAGs into continuous causal discovery

- Problem formulation
- Background: continuous DAG learning and plug-and-play priors
- DAG-PnP: plug-and-play causal discovery with diffusion priors
- Theoretical guarantees
- Results

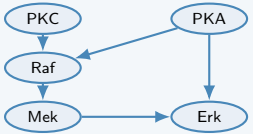
- Data $\mathbf{X} \in \mathbb{R}^{n \times d}$ (n samples)
- Linear SEM

$$\mathbf{X} = \mathbf{X}\mathbf{W} + \mathbf{E}$$

$\Rightarrow W_{ij} \neq 0 \iff$ effect $i \to j$

$\Rightarrow \mathcal{G}(\mathbf{W})$ must be a **DAG**

Example: protein signaling



Raf is generated from its parents:

$$x_{\mathrm{Raf}} = W_{\mathrm{PKC,Raf}}\, x_{\mathrm{PKC}} + W_{\mathrm{PKA,Raf}}\, x_{\mathrm{PKA}} + e_{\mathrm{Raf}}$$

- **Goal**: estimate a DAG-compatible $\mathbf{W}$

$$\min_{\mathbf{W}} \ \mathcal{L}_{\mathrm{data}}(\mathbf{W}) \quad \text{s.t.} \quad \mathcal{G}(\mathbf{W}) \in \mathbb{D} \quad \text{(set of DAGs)}$$

- Data fidelity $\mathcal{L}_{\mathrm{data}}$: ordinary least-squares loss

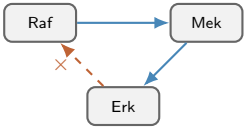
$$\mathcal{L}_{\mathrm{data}}(\mathbf{W}) = \frac{1}{2n} \|\mathbf{X} - \mathbf{X}\mathbf{W}\|_F^2$$

$\Rightarrow$ Noise model: $E_{ij} \sim \mathcal{N}(0, \sigma^2)$ i.i.d., **equal variances** across variables

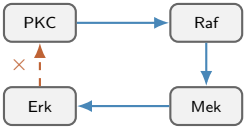
- **Observational data**: passive measurements, no interventions, few samples
- **Direction is ambiguous**: one dependence, several causal explanations
- **Acyclicity**: combinatorial and global $\rightarrow |\mathbb{D}_8| \approx 7.8 \times 10^{11}$ DAGs on 8 nodes



(a) both directions fit



(b) no cycles allowed



(c) global: one edge closes a long cycle

- Need structural priors beyond data fit, while keeping computation in check

- Score-based search over DAGs $\min_{\mathcal{G}(\mathbf{W})\in\mathbb{D}} \mathcal{S}(\mathbf{W};\mathbf{X})$ is combinatorial
- NOTEARS [Zheng 2018]: smooth function h whose zero level set is exactly $\mathbb{D}$

$$\min_{\mathbf{W}} \mathcal{S}(\mathbf{W};\mathbf{X}) \quad \text{s.t.} \quad h(\mathbf{W}) = 0 \quad \rightarrow \text{first-order methods}$$

  $\Rightarrow$ NOTEARS: $h(\mathbf{W}) = \operatorname{tr}(e^{\mathbf{W}\circ\mathbf{W}}) - d$ (trace counts cycles; $-d$ removes the identity)
- DAGMA [Bello 2022]: log-determinant characterization, better detects long cycles

$$h(\mathbf{W}) = -\log\det(s\mathbf{I} - \mathbf{W}\circ\mathbf{W}) + d\log s$$

  $\Rightarrow$ Loop view: $h(\mathbf{W}) = \sum_{k\geq 2} \operatorname{tr}\big((\mathbf{W}\circ\mathbf{W})^k\big)/(k\, s^k) \rightarrow$ larger s: more room, weaker push
  $\Rightarrow$ Defined on $\mathcal{W}_s = \{\mathbf{W} : \det(s\mathbf{I} - \mathbf{W}\circ\mathbf{W}) > 0\}$, and $h(\mathbf{W}) = 0 \iff \mathcal{G}(\mathbf{W})$ is a DAG
  $\Rightarrow$ Gradient $\nabla h(\mathbf{W}) = 2\mathbf{W}\circ(s\mathbf{I} - \mathbf{W}\circ\mathbf{W})^{-\top}$ blows up near the boundary $\partial\mathcal{W}_s$
- Limitation: only data fit + acyclicity, regularization stays **hand-crafted** (e.g., ℓ_1)
  $\Rightarrow$ Simple graph structures
  $\Rightarrow$ Propose to use learned denoiser as implicit priors. How?

- Regularized (MAP-type) estimation: $\min_{\mathbf{W}} \; F(\mathbf{W}) + \mathcal{R}(\mathbf{W})$
- Proximal gradient alternates two steps

$$\mathbf{Z} = \mathbf{W}^{(k)} - \gamma \nabla F(\mathbf{W}^{(k)}) \qquad \text{gradient step}$$

$$\mathbf{W}^{(k+1)} = \operatorname*{argmin}_{\mathbf{W}} \; \frac{1}{2\sigma^2} \|\mathbf{W} - \mathbf{Z}\|_F^2 + \mathcal{R}(\mathbf{W}) \qquad \text{proximal step}$$

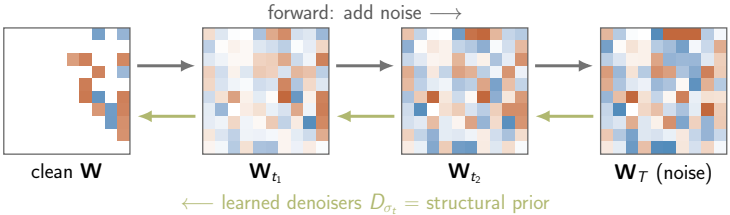
  $\Rightarrow$ The proximal step is **MAP denoising** of $\mathbf{Z}$ at noise level σ
- Plug-and-play [Venkatakrishnan 2013]: replace the proximal step with **any denoiser** $D_\sigma(\mathbf{Z})$
  $\Rightarrow$ The prior is **implicit**: $\mathcal{R}$ is never needed in closed form
  $\Rightarrow$ Separates the data-fidelity objective from the prior $\rightarrow$ modular
- Our idea: use a diffusion model trained on DAGs as the denoiser

- Forward process: gradually corrupt a clean DAG $\mathbf{W} \sim p(\mathbf{W})$ with Gaussian noise

$$\mathbf{W}_t = \alpha_t \mathbf{W} + \sigma_t \boldsymbol{\epsilon}, \qquad [\boldsymbol{\epsilon}]_{ij} \sim \mathcal{N}(0,1)$$

- Reverse process: learn the score $\nabla \log p_t(\mathbf{W}_t)$ to iteratively denoise
- Tweedie's formula $\rightarrow$ multi-scale family of denoisers $\{D_{\sigma_t}\}_{t=1}^{T}$

$$D_{\sigma_t}(\mathbf{W}_t) = \mathbb{E}[\mathbf{W} \mid \mathbf{W}_t] = \frac{1}{\alpha_t}\left(\mathbf{W}_t + \sigma_t^2 \nabla \log p_t(\mathbf{W}_t)\right)$$



- Trained on weighted DAGs (score matching [Jo 2022]) with a **direction-aware** dual-attention network [Carballo 2025]

	Learned prior	Designed for DAGs	Efficient
Continuous optimization NOTEARS, DAGMA	×	✓	✓
Full posterior over DAGs e.g., DAG-GFlowNet	hand-specified	✓	×
Diffusion for graph inverse problems PRODIGY, ...	✓	× (undirected)	guided sampling
This work: DAG-PnP (MAP estimate)	✓	✓	✓

- Contributions
  - ⇒ A diffusion prior over weighted DAGs, plugged into a gradient-based method
  - ⇒ A barrier-aware step size that preserves feasibility near $\partial \mathcal{W}_s$
  - ⇒ A local convergence guarantee for the resulting fixed-point iteration

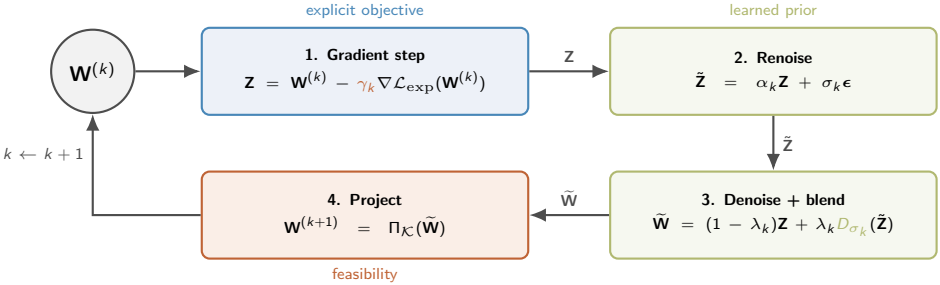
- Frame causal discovery as

$$\min_{\mathbf{W}\in\mathcal{W}_s} \underbrace{\mathcal{L}_{\text{exp}}(\mathbf{W})}_{\text{explicit}} + \underbrace{\mathcal{R}_{\text{prior}}(\mathbf{W})}_{\text{implicit (learned)}}$$

- Explicit objective: data fit + acyclicity

$$\mathcal{L}_{\text{exp}}(\mathbf{W}) = \lambda_{\text{data}}\,\mathcal{L}_{\text{data}}(\mathbf{W}) + \lambda_{\text{dag}}\,h(\mathbf{W})$$

- Implicit regularizer $\mathcal{R}_{\text{prior}}$: plays the role of $-\log p(\mathbf{W})$
  - $\Rightarrow$ Accessed only through the diffusion denoiser D_σ
  - $\Rightarrow$ No ℓ_1 penalty needed: the prior is trained on sparse DAGs $\rightarrow$ sparsity is implicit
- Three ingredients
  - $\Rightarrow$ Gradient step on $\mathcal{L}_{\text{exp}}$ · Denoising step with D_σ · Feasibility (step size + projection)



- Gradient: enforces agreement with the data and the acyclicity penalty
- Renoise + denoise: D_{σ_k} expects an input at noise level $\sigma_k \rightarrow$ forward-diffuse $\mathbf{Z}$ first
  $\Rightarrow \lambda_k \in [0,1]$ controls how strongly the prior acts at iteration k
- Project: the denoiser is not guaranteed to stay inside $\mathcal{W}_s$

- Noise levels inherited from the variance-preserving (VP) schedule of the diffusion model

$$\alpha_k = \sqrt{\bar{\alpha}_k}, \qquad \sigma_k = \sqrt{1 - \bar{\alpha}_k}, \qquad \bar{\alpha}_k = \prod_{i=0}^{k} (1 - \beta_i)$$

  $\Rightarrow$ DAG-PnP traverses the schedule in reverse: from high noise to low noise

- Blend weight $\propto$ inverse signal-to-noise ratio [Mardani 2023]

$$\lambda_k = \frac{\sigma_k}{\alpha_k}$$

  $\Rightarrow$ Early: high noise, strong prior $\quad\rightarrow\quad$ Late: data fit dominates

- Projection onto a compact safe set

$$\mathcal{K} = \left\{ \mathbf{W} \in \mathcal{W}_s : s_{\min}(s\mathbf{I} - \mathbf{W} \circ \mathbf{W}) \geq \delta_{\min},\ \|\mathbf{W}\|_F \leq W_{\max} \right\}$$

  $\Rightarrow$ Keeps iterates strictly inside $\mathcal{W}_s \rightarrow h$ and ∇h stay finite

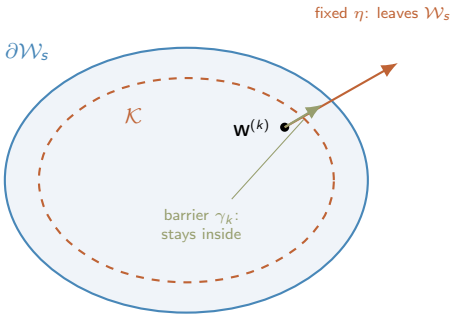
- Distance to the boundary: smallest singular value

$$\delta(\mathbf{W}) = s_{\min}(s\mathbf{I} - \mathbf{W} \circ \mathbf{W})$$

- ∇h contains $(s\mathbf{I} - \mathbf{W} \circ \mathbf{W})^{-1}$, so it grows as $O(1/\delta)$ near $\partial\mathcal{W}_s$
  - $\Rightarrow$ A fixed step size can leave the domain
- Our interior-point-style rule

$$\gamma_k = \min\left(\eta,\ c\,\frac{\delta(\mathbf{W}^{(k)})}{\|\nabla\mathcal{L}_{\exp}(\mathbf{W}^{(k)})\|_F + \epsilon}\right)$$

  - $\Rightarrow$ η: base step, $c \in (0,1)$: safety factor
  - $\Rightarrow$ ϵ: avoids division by zero at stationary points



$$\gamma_k = \min\left(\eta,\ c\,\frac{\delta(\mathbf{W}^{(k)})}{\|\nabla\mathcal{L}_{\exp}(\mathbf{W}^{(k)})\|_F + \epsilon}\right)$$

1. Far from the boundary

- Moderate $\nabla\mathcal{L}_{\exp}$, δ not small
- min selects η
- Standard gradient step

2. Near a stationary point

- $\nabla\mathcal{L}_{\exp} \to \mathbf{0}$
- ϵ keeps the ratio finite, min selects η
- Step does not blow up

3. Near the boundary

- $\delta \to 0$, $\|\nabla\mathcal{L}_{\exp}\|_F$ huge
- Move: $\gamma_k\|\nabla\mathcal{L}_{\exp}\|_F \le c\,\delta$
- Gradient size cancels $\to$ stays inside $\mathcal{W}_s$

Require: Data $\mathbf{X}$, denoiser D_σ, schedule $\{\alpha_k, \sigma_k, \lambda_k\}_{k=0}^{K-1}$

1: Initialize $\mathbf{W}^{(0)} \in \mathcal{K}$
2: **for** $k = 0, 1, \ldots, K-1$ **do**
3: Compute $\nabla \mathcal{L}_{\exp}(\mathbf{W}^{(k)})$ and γ_k
4: $\mathbf{Z} = \mathbf{W}^{(k)} - \gamma_k \nabla \mathcal{L}_{\exp}(\mathbf{W}^{(k)})$ ▷ Gradient
5: $\tilde{\mathbf{Z}} = \alpha_k \mathbf{Z} + \sigma_k \boldsymbol{\epsilon}, \ \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ ▷ Renoise
6: $\tilde{\mathbf{W}} = (1 - \lambda_k)\mathbf{Z} + \lambda_k D_{\sigma_k}(\tilde{\mathbf{Z}})$ ▷ Denoise
7: $\mathbf{W}^{(k+1)} = \Pi_{\mathcal{K}}(\tilde{\mathbf{W}})$ ▷ Project
8: **end for**
9: **return** $\mathbf{W}^{(K)}$

- Same cost per iteration as DAGMA + one denoiser call
- DAGMA = DAG-PnP with identity denoiser $D_\sigma(\mathbf{W}) \equiv \mathbf{W}$
- Prior is modular: swap the denoiser for another graph family

- Deterministic simplification (fixed σ, $\lambda_k = 1$, no renoising):
  $T_\sigma(\mathbf{W}) = \Pi_{\mathcal{K}}\big(D_\sigma(\mathbf{W} - \gamma(\mathbf{W})\nabla\mathcal{L}_{\exp}(\mathbf{W}))\big)$

Lemma 1 (Feasibility)

For a small enough safety factor c, the barrier-aware gradient step never leaves the DAGMA domain: if $\mathbf{W} \in \mathcal{K}$, then $\mathbf{W} - \gamma\nabla\mathcal{L}_{\exp}(\mathbf{W}) \in \mathcal{W}_s$.

Theorem 1 (Local convergence)

If T_σ is a contraction (factor $\rho < 1$) on a neighborhood $\mathcal{U}$ of a solution, and the iterates enter $\mathcal{U}$, they converge to a unique fixed point $\mathbf{W}^\star$ at a linear rate: $\|\mathbf{W}^{(k)} - \mathbf{W}^\star\|_F \le \rho^k \|\mathbf{W}^{(0)} - \mathbf{W}^\star\|_F$.

- Proof ideas: each step is shorter than the distance δ to the boundary; Banach fixed point
- $\mathbf{W}^\star$ balances data fit, acyclicity, and the learned prior
- Local: nonconvex problem and a learned denoiser; the full stochastic algorithm remains open

- Synthetic linear SEMs $\mathbf{X} = \mathbf{X}\mathbf{W} + \mathbf{E}$ with Gaussian noise
  - $\Rightarrow$ Erdős–Rényi (ER4) and scale-free (SF4) graphs, $\approx 4d$ edges
  - $\Rightarrow$ **Low-data regime**: $n = d$, with $d \in \{20, 50\}$; 10 random seeds
- Prior: diffusion model trained on ER4 and SF4 DAGs with $d = 20$
- Baselines
  - $\Rightarrow$ NOTEARS [Zheng 2018], DAGMA [Bello 2022] (= DAG-PnP without denoiser), Greedy-BIC
- Metrics
  - $\Rightarrow$ Structural Hamming distance (SHD$\downarrow$) and F1$\uparrow$ on directed edges
  - $\Rightarrow$ Acyclicity violation $h(\hat{\mathbf{W}}) \downarrow$ and fit error $\mathcal{L}_{\text{data}}(\hat{\mathbf{W}}) \downarrow$

		ER4				SF4				
d	Method	SHD↓	h ↓	Fit↓	F1↑	SHD↓	h ↓	Fit↓	F1↑	Time (s)
20	Greedy-BIC	**89.3**	**0.000**	8.57	0.244	**88.5**	**0.000**	8.41	0.206	1.4
	NOTEARS	142.8	0.239	1.09	0.359	145.1	0.216	1.19	0.326	0.6
	DAGMA	146.5	0.585	**0.77**	0.291	147.1	0.535	**0.87**	0.269	4.1
	DAG-PnP	130.2	0.176	1.43	**0.391**	135.2	0.150	1.57	**0.345**	6.1
50	Greedy-BIC	**255.5**	**0.000**	23.96	**0.268**	**256.1**	**0.000**	23.03	**0.242**	101.0
	NOTEARS	888.6	0.215	3.41	0.230	895.1	0.209	3.66	0.207	1.6
	DAGMA	931.6	0.487	**2.25**	0.213	921.5	0.499	**2.43**	0.199	1.8
	DAG-PnP	820.2	0.213	3.61	0.246	806.4	0.195	3.93	0.229	6.0

Bold: best. Underline: second best. Mean over 10 seeds.

- Best SHD and h among continuous methods; best F1 overall at $d = 20$
- Higher fit error: the prior trades data fit for plausible structure
- Greedy-BIC wins SHD, but its runtime grows $\sim 72\times$ from $d = 20$ to $d = 50$

  $\Rightarrow$ DAG-PnP stays at $\approx 6\,\mathrm{s}$ for both sizes

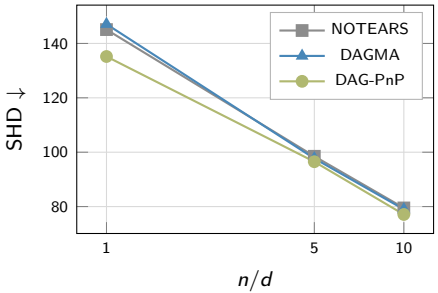
- Isolate the two feasibility mechanisms at $d = 20$ (SF4)

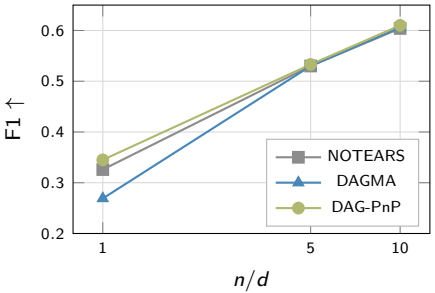
Configuration	SHD↓	h ↓	Fit↓	F1↑
Fixed step, no projection	162.9	5.30	**0.002**	0.300
Barrier step, no projection	161.7	5.39	**0.002**	0.298
Fixed step, projection	139.0	0.23	1.33	0.326
Barrier step, projection	**135.2**	**0.15**	1.57	**0.345**

- Without projection: large acyclicity violations for both step sizes (away from the feasible set)
- Projection alone substantially improves structure recovery
- Barrier step + projection: lowest SHD and h, highest F1

  $\Rightarrow$ Slightly higher fit error: trades data fit for acyclicity

- Fix $d = 20$ (SF4) and vary $n \in \{d, 5d, 10d\}$





- Largest gain in the **low-data regime** $n/d = 1$

  $\Rightarrow$ The prior compensates for data scarcity

- As n grows, the data become informative and all methods perform similarly

- Protein signaling data [Sachs 2005]: $d = 11$ proteins, $n = 853$ cells
  - $\Rightarrow$ Partially known ground-truth DAG with 17 edges
  - $\Rightarrow$ Single ground truth $\rightarrow$ we reuse the synthetic ER4 prior

Method	SHD↓	$h \downarrow$	Fit↓	F1↑
Greedy-BIC	**23.0**	**0.000**	1.91	0.26
NOTEARS	44.9	0.002	**0.98**	0.31
DAGMA	29.4	0.015	2.19	0.30
DAG-PnP	26.7	0.019	5.44	**0.45**

- DAG-PnP: highest F1 and lowest SHD among continuous methods

- We proposed DAG-PnP, a plug-and-play method for causal discovery
  - $\Rightarrow$ Combines SEM data fit + DAGMA acyclicity with a diffusion prior over weighted DAGs
  - $\Rightarrow$ Barrier-aware step size + projection keep the iterates feasible
  - $\Rightarrow$ Local fixed-point convergence for the deterministic iteration
  - $\Rightarrow$ The learned prior helps most when **data are scarce**
- Limitations
  - $\Rightarrow$ Linear SEMs only; local theory for the fixed-σ variant
  - $\Rightarrow$ The diffusion model must be trained on a graph family relevant to the problem
- Future work
  - $\Rightarrow$ Nonlinear SEMs
  - $\Rightarrow$ Denoisers that map into $\mathcal{W}_s$ by construction
  - $\Rightarrow$ Global convergence guarantees

Code: `github.com/aliaaz99/dag-pnp`

- Shared across methods: $\lambda_{\text{data}} = 1.0$, $\lambda_{\text{dag}} = 0.01$, $\eta = 0.1$, $K = 1000$ iterations
- Barrier step: $c = 0.3$, $\epsilon = 10^{-6}$, $\delta_{\min} = 10^{-6}$
- VP schedule: $\beta_{\min} = 0.1$, $\beta_{\max} = 1.0$
- DAGMA parameter: $s = 1.5\,(2k\, w_{\max}^2) + 1$
  - $\Rightarrow$ From the Gershgorin bound $\rho(\mathbf{W} \circ \mathbf{W}) \leq 2k\, w_{\max}^2$, independent of d for ERk/SFk
- Projection radius: $W_{\max}(d) = 3\sqrt{kd\,\overline{w^2}} + 1$
  - $\Rightarrow$ Avoids clipping the true graph, since $\mathbb{E}\left[\|\mathbf{W}_{\text{true}}\|_F^2\right] = kd\,\overline{w^2}$
- Greedy-BIC: adds the edge with the largest local BIC gain

$$\text{score}(j \mid \text{Pa}(j)) = -\frac{n}{2} \log \frac{\text{RSS}_j}{n} - \frac{|\text{Pa}(j)| \log n}{2}$$