Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Audited Benchmark for Minimum Weighted Feedback Arc Set: Graph Identity, Algorithm Comparison, and Reproducibility

Soroush Vahidi

sv96@njit.edu

New Jersey Institute of Technology

Research Article

Keywords: minimum weighted feedback arc set, benchmark dataset, heuristic comparison, numerical optimality bounds, reproducibility

Posted Date: October 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11436856/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

An Audited Benchmark for Minimum Weighted Feedback Arc Set: Graph Identity, Algorithm Comparison, and Reproducibility

Soroush Vahidi[1*]

[1*]New Jersey Institute of Technology, Newark, New Jersey, USA.
ORCID: https://orcid.org/0000-0003-1934-6282.

Corresponding author(s). E-mail(s): sv96@njit.edu;

Abstract

Dependable comparisons of minimum weighted feedback arc set (MWFAS) heuristics require consistent inputs, computational budgets, and treatment of incomplete runs. We present an audited benchmark linking weighted graphs and source-parent relationships to validated saved orders, original outcomes, and numerical optimization references. This resource extends an earlier metrics-only release and supports comparisons across graph families, paired weight regimes, and empirical sources. Incumbent-protected strongly connected component neighborhood search (IPSNS) showed family-dependent gains over its better initializer. Across all tested graph–seed groups in two generated-graph campaigns, ten iterations produced the same saved objectives and permutations as every larger tested budget. In the replicated campaign, family median process-wall ratios relative to the selected initializer were 1.27–1.60 at ten iterations, compared with 3.19–9.82 at 200. This observed trade-off is not a convergence or stopping guarantee. Under matched wall-clock horizons, IPSNS seed 1 beat RASstar on seven of nine selected synthetic topologies but only one of four selected empirical parents. At 300 s, single-node insertion from either evaluated initializer beat IPSNS seed 1 on all four parents. In a separate empirical campaign, complete 200-iteration comparisons covered only 18 of 38 views, all from one source parent. These findings show how budget, availability, dependence, and comparator choice affect interpretation. Public saved-output replay covers all ten numerical tables, Figure 2, and 11/17 tracked claim groups; six broader groups and historical-source reconstruction remain unresolved. The evidence supports scoped comparisons of existing methods rather than a universal ranking.

Keywords: minimum weighted feedback arc set, benchmark dataset, heuristic comparison, numerical optimality bounds, reproducibility

Acknowledgments

The author thanks his PhD advisor, Professor Ioannis Koutis, for academic guidance and support; Anders Borum, developer of Secure ShellFish, for complimentary lifetime access supporting remote research and development; his friend Artin Majd for emotional support and encouragement; and his mother for emotional support, understanding, and encouragement. He acknowledges the computational resources of New Jersey Institute of Technology's Wulver high-performance computing cluster.

1 Introduction

1.1 Motivation and problem context

Directed comparisons need not be transitive. Pairwise preferences can contain cycles, as can communication and input–output networks once their arcs are interpreted as directional constraints. After removing source self-loops, an order of a directed graph incurs the total weight of arcs that point backward. Deleting those arcs leaves the loop-free graph acyclic. This order-based view of minimum weighted feedback arc set (MWFAS) connects to linear ordering and ranking [19, 30, 44], while directed feedback algorithms offer other ways to construct solutions [11].

Exact mixed-integer optimization supplies certified numerical optima and lower bounds, but its cost can rise sharply on some graphs; a time limit may leave a feasible order and an unresolved gap. Heuristics remain relevant when many instances or larger networks must be processed. Classical greedy ordering and recent scalable methods illustrate the algorithmic range [2, 14, 47]. Their existence does not establish which method works well when family, weights, or strongly connected structure differ from those in its original evaluation.

1.2 Limits of benchmark interpretation

An average objective can hide reversals between graph families, and raw backward weight scales with graph size and arc weights. A certified optimum differs from a feasible incumbent with a time-limited lower bound. Runtime includes conversion and initialization; a capped run has an unobserved completion time. Seeds and derived views of one source are dependent observations. Linear-ordering benchmarks and algorithm-assessment research motivate explicit metrics, budgets, and statistical units [12, 31, 36, 45].

The original MWFAS Heuristic Metrics release contains 7,371 historical metric rows in 14 configurations [43], but lacks the graph and ordering bytes needed to rescore every row. Our extension links generated and empirical inputs to saved solutions, numerical references, and original outcomes. This enables comparisons that distinguish shared inputs from independent evidence and preserve the effects of missing results and computational budgets.

1.3 Research questions and contributions

RQ1: Paired-input algorithm comparison. How do the evaluated MWFAS implementations differ in validated objective quality on the same weighted inputs across tested graph families and source parents, and what changes under a common wall-clock budget? **RQ2: Benchmark variation.** How do graph topology, weights, and provenance change interpretation of those comparisons, and which structural associations remain exploratory? **RQ3: Reference quality.** How close are validated heuristic orders to solver-certified numerical optima or validated lower and feasible upper bounds? **RQ4: Reliability and reuse.** How do budgets, failures, censoring, and artifact access constrain comparative and reproducibility claims?

The study makes two contributions: an original, reusable empirical resource linking graph identities and source parents to validated saved solutions and preserved outcomes; and scoped evidence of how budgets, weighting, dependence, availability, and comparator choice affect heuristic comparisons. The results distinguish refinement quality from its computational cost, reveal source concentration among complete cases, and show changes in preferred method across cohorts and comparators. Most comparisons pair inputs under unequal computation; a separate selected cohort matches end-to-end wall-clock horizons. The contribution is a benchmark and comparison of existing methods, distinct from the author's earlier IPSNS work and published ranking pipeline [44].

Sections 3–4 define the metrics and protocols, Section 5 describes the benchmark, and Sections 6–9 evaluate quality, reference gaps, structure, and reliability. Section 10 brings these findings together and discusses their practical scope.

2 Related work

2.1 Classical feedback-arc-set algorithms

The Eades–Lin–Smyth heuristic builds an order from local degree information and remains a useful historical reference [14]. Combinatorial feedback algorithms demonstrate the value of graph reductions and approximation-oriented reasoning [11]. Ranking approximation results provide guarantees for particular input models; those guarantees do not automatically extend to arbitrary directed weights [26]. General directed weighted feedback-edge deletion has a separate approximation literature [16]; none of those theoretical guarantees is attributed here to the particular frozen heuristic implementations. The frozen suite also evaluates LR-TA, a local-ratio-based ordering implementation, and WMSF/L2. Their exact semantics come from pinned source revisions, rather than an assumed equivalence to every method sharing an algorithmic label.

2.2 Modern scalable heuristics

Xiong and colleagues combine optimum-preserving graph reductions with a scalable divide-and-conquer feedback-arc heuristic [47]. Their FASDC algorithm interleaves reduction, SCC decomposition, and balanced partitioning of greedy orders; its preferred partition ratio varies by graph family in their experiments. Their main large-graph evaluation concerns arc cardinality, while a separate test assigns positive integer weights to the source arcs. Our source audit identifies the authors' `RASstar` executable as the FASDC branch with $r = 1/3$. We evaluate that integrated configuration on unchanged weighted inputs and independently rescore its saved orders, without isolating the reduction rules or testing the full algorithm portfolio. These protocol and weight differences matter when interpreting its results alongside ours. Rocket-Crane combines continuous weighted ordering with later refinement stages [2]; an audited C component is not its full pipeline. Hecht and colleagues' TIGHT-CUT work includes weighted feedback-set theory [21]. Cavallaro and colleagues study minimal and stable feedback sets through orderings and centrality-based refinements [8, 9], with a separate weighted method report [7]. Cardinality, minimality, stability, and weighted backward cost are distinct targets. The external Rocket-Crane, TIGHT-CUT, and Cavallaro implementations were not authenticated as compatible full-order competitors in these tables. Evaluated WMSF/L2 is the project's pinned implementation, not an original-author Cavallaro executable. IPSNS and its earlier sparse-benchmark results were publicly described in Vahidi's 2026 SSRN preprint (https://doi.org/10.2139/ssrn.6281222); the present study treats IPSNS as an existing method. A separate Vahidi–Koutis preprint studies interval refinement on the FlyWire connectome (https://doi.org/10.2139/ssrn.6221201); neither its method nor its FlyWire results are evaluated here.

2.3 Exact optimization and linear ordering

Pairwise precedence variables and transitivity constraints connect MWFAS to the linear ordering problem (LOP). Cutting-plane and benchmark work establish exact and heuristic reference points [19, 30, 31]. Ben Ali's CL-TLBO searches populations of permutations: exemplar precedences guide recursive pivot partitioning, and restricted single-item insertion refines subproblems at multiple levels [4]. Qian and colleagues show how moving a *block* can improve an order already locally optimal under single-node insertion. Their block-insertion search feeds iterated and memetic methods, with an additional parallel exact-subproblem framework [37]. Qian and colleagues compare randomized methods under wall limits calibrated from a nominal $1000n^2$ fitness-evaluation budget; Ben Ali also reports both wall-time and fitness-evaluation comparisons on established LOP families. Both works maximize pairwise forward benefit on LOP matrices, using protocols that differ from our iteration and process-wall caps. For nonnegative off-diagonal weights, maximizing forward matrix weight is equivalent to minimizing backward graph weight only after direction, loop, parallel-arc, and declared-vertex conventions agree. No authenticated original-author implementation of either method was executed here. In particular, the evaluated project routine moves single nodes from a fixed initializer; it is not Qian's block-insertion

algorithm or Ben Ali's population search. These methods therefore remain prospective comparators. Lugo and colleagues' MA-EDM is a diversity-aware memetic LOP method combining population search with local improvement [29]. Pérez Lugo and colleagues' later work examines stronger intensification as computational budgets increase [35]. These contributions make population management and search effort relevant comparator dimensions; their published results are not same-input observations here. The separate EXP31 feasibility pilot used the 2022 implementation, not the 2025 successor, and is not a homogeneous, validated performance comparison. Ando and colleagues' 2026 preprint (arXiv:2606.24558) reports that insertion search narrows differences among tested constructive LOP heuristics on xLOLIB; its methods and instances are not a same-input baseline here. The dense LOP literature informs comparator choice; its published outcomes are not pooled with our sparse-graph results. Benito-Marimon and colleagues study multiple optimal LOP orders and their ranking diversity [5]; a tie in backward cost need not imply identical orders or application utility. Schiavinotto and Stützle examined LOP search spaces and metaheuristics [38]; Fagiolo and colleagues discuss economic LOP structure [17]. Baharev and colleagues' exact cycle-enumeration method [3] was not used in our HiGHS or Gurobi protocols. A solver may yield an incumbent, a lower bound, or a certified numerical optimum; the audited exact subset contains both certified and time-limited cases.

2.4 Algorithm benchmarking and instance difficulty

Anytime benchmarking asks how outcomes evolve with budget, rather than considering a final objective alone [45]. Algorithm-selection work highlights metric, preprocessing, and test-design sensitivity [36]. A parametrized scheduling benchmark uses controlled instance families and exact optimization to distinguish structural constraints, attainable objective quality, and solver effort [18]; its scheduling-specific indicators are not measures of MWFAS difficulty. Statistical comparisons across datasets require genuine independent units [12]: generated topology is the principal unit here, while seeds and paired weights are repeated observations. Software-artifact research emphasizes versioned code and explicit availability statements for reproducible analysis [25]; here a public metadata table alone cannot rescore an unpublished graph/order pair. Performance profiles provide an established way to compare optimization software [13], but a pooled profile would be misleading when candidate availability, clocks, stopping conditions, and censoring are not aligned; we therefore report protocol-specific paired quality and completion status instead. In real-world analysis, derivative views from four source parents cannot become 38 independent datasets. Strongly connected components (SCCs) are measured with standard directed-graph algorithms [40]; SCC size is a structural descriptor, not automatically heuristic error or solver effort.

2.5 Positioning of the present study

The earlier IPSNS preprint introduced refinement and sparse-benchmark results; the published ranking article studied acyclic construction and ranking scores. Here, the contribution is a reusable empirical resource and the comparisons enabled by linking exact inputs, source relationships, saved solutions, budgets, and outcomes. Hashing, provenance tracking, and statistical procedures serve that purpose as established methods.

The examined public MWFAS summaries lack saved per-instance orders and verified content joins needed for the comparisons reported here. Dense LOLIB matrices are not established matches to this corpus. LOLIB, SNAP, and the Ali collection supply established inputs and source documentation; the present resource adds the study-specific links among validated solutions, numerical references, original failures, parentage, and access conditions. This distinction concerns the examined releases, not global uniqueness. Reused inputs and algorithms remain explicitly attributed.

3 Problem formulation and evaluation metrics

3.1 Weighted directed graph representation

Let $G^{\mathrm{src}} = (V, A^{\mathrm{src}}, w^{\mathrm{src}})$ denote a directed source graph with a declared vertex set, finite nonnegative arc weights, and any source self-loops. Source

events with the same ordered endpoints are first aggregated by the documented input rule. Define the *optimization graph* $G = (V, A, w)$ by retaining only off-diagonal arcs, $A = \{(u, v) \in A^{\mathrm{src}} : u \neq v\}$, with their aggregated weights. The excluded loop weight is $C_{\mathrm{loop}} = \sum_{(v,v)\in A^{\mathrm{src}}} w^{\mathrm{src}}_{vv}$. Normalized experiment files store positive integer off-diagonal weights: zero matrix entries are omitted, negative values rejected, and isolated vertices retained. Graph identity includes declared vertices, orientation, and weights; a source identity additionally records its transformation and loop convention.

3.2 MWFAS and linear-ordering equivalence

A directed acyclic graph (DAG) contains no directed cycle. A complete order is a bijection $\pi : V \to \{1, \ldots, |V|\}$. On the loop-free optimization graph its backward weight is

$$B_G(\pi) = \sum_{(u,v)\in A} w_{uv} \mathbf{1}\{\pi(u) > \pi(v)\}. \quad (1)$$

Deleting these backward arcs *from* G leaves a DAG: every retained arc strictly advances in π. Conversely, if deleting $F \subseteq A$ makes G acyclic, any topological order of $G - F$ has backward arcs only in F and hence $B_G(\pi) \leq w(F)$ for nonnegative weights. The minimum of Eq. 1 therefore equals the minimum feedback-arc deletion weight on G; this is the ordering equivalence used here and in prior ranking work [44]. For the original source graph, every self-loop must also be deleted, so the corresponding deletion optimum is $C_{\mathrm{loop}} + \min_\pi B_G(\pi)$. An original-source order cost may likewise be reported as $C_{\mathrm{loop}} + B_G(\pi)$. Since C_{loop} is fixed for that source graph, it cannot change the ordering or ranking of algorithms evaluated on that same graph. These equivalences require complete orders, nonnegative weights, and the stated vertex, direction, and loop conventions.

For a nonnegative precedence-benefit matrix W whose off-diagonal entries use the same direction and weights as G, let $T_G = \sum_{i\neq j} W_{ij}$ and $F_G(\pi) = \sum_{\pi(i)<\pi(j)} W_{ij}$. Then $B_G(\pi) = T_G - F_G(\pi)$, including when some off-diagonal entries are zero. The EXIOBASE-derived instances use this linear-ordering transformation after excluding the matrix diagonal. Matrix orientation, integer scaling, diagonal treatment, and T_G must accompany every reported value.

3.3 Solution-quality metrics

Raw active-graph B_G is the auditable primary objective; any original-source total states C_{loop} separately. Cross-graph displays use $q_G = B_G/T_G$ when $T_G > 0$ and a paired difference $[B_G(\pi_a) - B_G(\pi_b)]/T_G$ on the same graph. This removes weight-scale differences but does not equalize intrinsic difficulty. When $T_G = 0$, every active order has zero cost and normalized values are undefined. Relative improvement against a reference order is $(B_{\mathrm{ref}} - B_{\mathrm{method}})/B_{\mathrm{ref}}$ only if $B_{\mathrm{ref}} > 0$; otherwise we report the absolute difference and tie status. Adding the same loop constant preserves strict wins and ties, but changes ratios with an original-source denominator; all reported normalized comparisons use the active graph. Strict wins, ties, and losses use independently recomputed objectives. The best of several seeds is an explicit selection rule on repeated runs, not several independent graph observations.

3.4 Exact bounds and optimality certification

Let H_g be a validated active-graph heuristic objective, U_g a validated feasible active-graph objective, and L_g a valid lower bound under the same objective convention. Then $L_g \leq \mathrm{OPT}_g \leq U_g$ and $H_g \geq \mathrm{OPT}_g$. A solver-optimal status with a checked complete order and bound agreement under documented tolerances supports a *solver-certified numerical* optimum, not an exact-arithmetic proof. A time-limited incumbent alone does not. If $\mathrm{OPT}_g > 0$, certified relative regret is $(H_g - \mathrm{OPT}_g)/\mathrm{OPT}_g$; at zero optimum, only an absolute difference is reported. With an open gap,

$$\max\{0, H_g - U_g\} \leq H_g - \mathrm{OPT}_g \leq H_g - L_g. \quad (2)$$

If $0 < L_g \leq U_g$, the corresponding relative interval is $[\max\{0, (H_g - U_g)/U_g\}, (H_g - L_g)/L_g]$; an upper relative endpoint is undefined when $L_g \leq 0$. The interval bounds one case's unknown regret, rather than statistical uncertainty. Adding C_{loop} to all active costs and bounds preserves absolute gaps, while changing relative-gap denominators; the reported gaps consistently exclude source

loops. Certified point gaps and bound intervals remain separate in the results.

4 Benchmark and experimental design

4.1 Benchmark sources and graph families

The generated cohorts span controlled families and sizes rather than a random sample of all directed networks. EXP03 has 45 directed graphs: 15 each from normalized Erdős–Rényi, scale-free, and community procedures, across three sizes and five target outdegrees. EXP04 independently generates 135 weighted topologies across three families, three sizes, three degrees, and five replicates per cell. The project generator begins with NetworkX directed $G(n, m)$, directed scale-free multigraph, or two-block stochastic-block draws [32]; it converts multigraphs to simple digraphs, removes loops, fixes $m = nk$, and repairs isolates as needed. Thus even the Erdős–Rényi-labelled outputs are not unmodified draws from a textbook random-graph law. EXP06 adds 99 topologies, including 18 project-defined planted-order cases, with two paired weight realizations each (198 weighted graphs). A planted hidden order is feasible, not presumed optimal. The independent SHA-256 join confirms 279 generated topologies across EXP03, EXP04, and EXP06.

EXP05 contributes 38 audited real-world graph views selected under a frozen configuration from four source parents. Eight sparse views derive from three SNAP parents: four count-weighted CollegeMsg views, two unit-weight Email-Eu-core views, and two unit-weight Wiki-Vote views. Thirty matrix-derived views share one EXIOBASE linear-ordering source. Windows, subgraphs, matrix variants, and other transformations retain their source-parent identity; a view, topology, weighted realization, and algorithm run are distinct units. Across EXP03–EXP06, 416 weighted graph hashes and 317 topology hashes are distinct, and their 5,083 original task records are repeated observations on those inputs. The 1,398 EXP06 records comprise 1,386 principal tasks (1,188 heuristic runs and 198 HiGHS solves) plus 12 separate igraph integer-programming cross-checks, not 1,398 graph instances. EXP07's 45 weighted graph hashes match nine EXP03 and 36 EXP04 inputs exactly; it adds no new independent graph. Neither its graph nor topology hashes overlap EXP06. EXP08's six-graph seed diagnostics inform sensitivity discussion only.

The later audited extensions add 20 present-day Ali Dasdan benchmark graph identities (EXP12), three source-parent-distinct empirical graphs from OpenFlights, Census LODES, and IRS SOI (EXP13), and three historical derived Wikimedia clickstream graphs (EXP14/15). Ali's 20 graph hashes are distinct within its collection, but its underlying independent source-parent count is unresolved. The Wikimedia files represent Arabic December 2025, Arabic January 2026, and Azerbaijani December 2025 views of one clickstream source parent; the pre-May-2026 monthly TSV snapshots have not been recovered. EXP10 reuses 18 EXP04 topologies and three EXP05 full views; EXP07 reuses 45 earlier graphs. The public EXP18 metadata projection indexes 442 graph rows under several normalization contracts, not 442 proven globally distinct weighted graphs. Its 5,254 outcome rows include repeated seeds, methods, and incomplete tasks; its 45 EXP07 reference records are separate from those outcomes. The original public dataset's 7,371 historical rows and 14 configurations remain unchanged and are not added to the extension outcome count. Historical row-to-input linkage is contextual for 1,525 of 1,607 name-linked Ali rows and unavailable for 82; none of these links proves historical input-byte identity.

Table 1 identifies the original providers and frozen source versions. The Ali collection documents mixed upstream origins; an Ali file name alone does not establish its original creator. Source-level detail, hashes, transformations, and rights status are recorded in the companion provenance register.

4.2 Instance preparation

Generated inputs record configuration, seed, complete vertex set, directed arcs, weights, and content hash. A separate topology hash excludes weights, allowing paired realizations to be recognized. EXP05's frozen selection and configuration identify source archives, transformations, parent IDs, graph and topology hashes, and DIMACS hashes. CollegeMsg parallel events become count

weights; Email-Eu-core and Wiki-Vote use unit weights. The audited EXIOBASE transformation excludes the matrix diagonal, omits zero entries, and preserves positive off-diagonal integer weights; all 30 source members and graph transformations passed independent checks. Sparse source self-loops are recorded as excluded source weight, including 642 on the full Email-Eu-core view; LODES and IRS similarly retain separate source-loop constants of 18,628 and 124,358,925. All reported method objectives optimize the loop-free graph. Two run IDs or graph-view names do not establish independent topologies.

4.3 Evaluated algorithms

Local-ratio topological add-back (LR-TA) and the weighted minimal/stable feedback-arc-set-style seed using its L2 ordering (WMSF/L2) produce deterministic initial orders in frozen project implementations. IPSNS constructs both complete seed orders, retains the lower-cost one, and then applies incumbent-protected strongly connected component neighborhood refinement; both construction costs are part of its full method runtime. Its seeds and iteration budgets are separate fresh runs, not checkpoints of one trajectory or independent graph observations. Weighted igraph/Eades returns a feedback-arc set; after its removal, a deterministic complete topological order is reconstructed and scored on the same loop-free graph. Its returned-set weight can differ from that order's backward weight; the latter is the common metric. EXP05 additionally uses project node-insertion refinement from each of LR-TA and WMSF, stopping at local optimality, a move limit, or a refinement time cap. EXP03 evaluates LR-TA, WMSF/L2 and IPSNS; EXP04/06 add weighted Eades; EXP05 adds insertion. EXP07 evaluates historical records against its exact evidence and adds no heuristic runs.

The pinned IPSNS implementation updates a copied incumbent only after a strict objective decrease and returns that snapshot. Thus a completed, validated run cannot worsen its selected initializer under the study's loop-free integer objective. This does not supply a missing order or validate a timeout checkpoint saved before both initializers finish; numerical or wrapper failures retain their original validation/status rules.

EXP10 invokes the original Xiong source as `RASstar` with $r = 1/3$, seed-dependent input shuffling, and the original positive arc weights; its weight-replacing option is omitted. The rank-by-vertex output is converted to a complete order and rescored on the unchanged weighted graph. EXP10 compares those saved orders against fresh 200-iteration IPSNS runs on reused inputs. EXP12/13 compare six method families: this RASstar configuration, LR-TA, WMSF/L2, both project insertion initializers, and weighted Eades; RASstar has two seeds per graph. EXP14 validates ten saved Wikimedia orders, but their method labels are historical context rather than a newly controlled contest. The exact commands, source revisions, seeds, and software environments are retained in the evidence ledger. Published-method lineage, project-maintained code, and externally authored executables are distinct provenance categories.

Table 2 separates methodological references from executed code. EXP28's restart arm repeatedly invokes the same RASstar configuration; its IPSNS arm instruments the pinned project implementation without changing search decisions. These are study protocols for existing algorithms.

4.4 Computational environments and budgets

The campaigns used frozen Wulver releases. EXP03's fresh IPSNS budgets were 10, 25, 50, 100, 200, and 400 iterations; EXP04 used 10, 50, 200, and 400; EXP05 used 1, 3, 10, 50, and 200 iterations with three seeds per budget and a 120-second per-task cap. Insertion additionally has at most 1,000 accepted moves and a 110-second refinement deadline checked between complete sweeps, within a separate 120-second subprocess cap. Initialization by LR-TA or WMSF and later refinement are separate costs in the recovered runs. An iteration label is not an equal-time budget. EXP06 used a project ordering MILP through the SciPy/HiGHS interface [39] with a 120-second solver cap and three fresh 200-iteration IPSNS seeds per weighted instance; EXP07 used Gurobi 12.0.3 [20] with project SCC decomposition, lazy transitivity cuts, and a 420-second cap with up to four threads.

Production manifests identify Wulver node `n0006` for EXP03–EXP06 and an AMD EPYC 7763 processor for EXP03–EXP05; the EXP06

Table 1 Dataset and generator provenance. A source parent may yield several dependent graph views. "Pinned" denotes the exact study input or revision, not a historical release date inferred from a file name. Full file hashes, release paths and rights status are in the companion provenance register.

Family / source	Provider / reference	Pinned identity	Study transformation	Use
Random-edge	$G(n, m)$ lineage [15]; NetworkX	Project seed, config and topology hash	Directed fixed-m draw; isolate repair; integer weights	03/04/06
Scale-free	Bollobás et al. [6]; NetworkX	Project seed, config and topology hash	Collapse parallel arcs; remove loops; fix m and isolates	03/04/06
Community	Holland et al. [22]; NetworkX	Project seed, config and topology hash	Directed block draw; remove loops; fix m and isolates	03/04/06
Planted order	Project-defined generator	EXP06 selection and topology hash	Sample pairs; reverse fixed fraction relative to hidden order	06
Paired weights	Project-defined sampling	EXP06 uniform/capped-lognormal seeds	Same topology; integer $U[1, 100]$ or capped lognormal weights	06
CollegeMsg	Panzarasa et al.; SNAP [27, 34]	Pinned SNAP temporal file	Temporal messages to count-weighted directed arcs/views	05/10/28
Email-Eu-core	Yin et al.; SNAP [27, 48]	Pinned SNAP static file	Unit arcs; loops excluded with recorded constant	05/10/28
Wiki-Vote	Leskovec et al.; SNAP [27, 28]	Pinned SNAP vote file	Unit directed arcs; full and SCC views	05/10
EXIOBASE LOP	Fagiolo et al. [17]; 2F/Zenodo [1]	Zenodo 20554376, v0.0.1, `exiobase.zip`	Positive off-diagonal matrix entries to weighted arcs	05/28
Ali/Dasdan	Dasdan collection [10]	Git revision `77e5b8b4`; 20 selected files	DIMACS arc weight; graph-byte identities; mixed upstream origins	12
OpenFlights	OpenFlights [33]	Git revision `56a3cce4`; routes/airports	U.S. nonstop routes; parallel record counts, not flights	13/28
LODES	U.S. Census Bureau [41]	LODES8 DE 2022, `de_od_main_JT00`	Block flows to directed tract job counts; loops separate	13
IRS SOI	Internal Revenue Service [24]	2021–22 county outflow CSV	County outflow return counts; suppressed/aggregate rows excluded	13
Clickstream	Wikimedia Foundation [46]	Three pinned derived graphs; old TSV unavailable	Link counts, historical threshold and label mapping	14/15
Original metrics	Vahidi [43]	HF revision `5b6cd562`	Historical metric rows; graph/order bytes partly absent	01/12
Benchmark extension	Vahidi [42]	Versioned HF releases; Data availability	Metadata, saved outputs, exact upstream transformations	03–07, 12–15, 28

manifest does not record a CPU model. Inputs, formulations, limits, and process clocks differ, so EXP06 and EXP07 are not paired solver-speed experiments. Exact software, resource, source, and configuration identifiers are retained in the evidence ledger. Local EXP08 timing is not directly comparable to Wulver timing. EXP10 used two fresh seeds per method on 18 selected EXP04 graphs and three EXP05 full views with a 120-second task cap; the source RASstar and IPSNS executions did unequal work. EXP12/13 used a common 120-second process cap per task, which does not equalize actual work. Their process wall and CPU records must remain distinct from EXP03–07 clocks and from EXP14's historically unverified timing labels.

EXP11 attempted a separate controlled-wall-cap pilot on five reused graphs. Of its 78 tasks, 37 passed, 19 were valid timeouts, and 22 were invalid under the strict end-to-end cap: 56 task statuses passed independent audit, but only 51 tasks supplied usable objectives. Separately audited EXP22 recoveries validate 16 of those originally invalid task identities, giving 72 valid statuses and 61 usable objectives across the same 78 identities; six one-second cases remain invalid. These pilot records are not included in the extension outcome count or the numerical tables and do not establish a comprehensive equal-time comparison.

The separate EXP28 controlled-wall study reuses nine selected generated topologies (three per family, from EXP04/06) and one view from each of four empirical source parents (CollegeMsg, Email-Eu-core, EXIOBASE, and OpenFlights); it adds no independent graph sample. Its 117 tasks are two project-maintained IPSNS seeds

Table 2 Algorithm and implementation provenance. Versions identify the tested code or library, not a claim that a named paper's executable was replicated. Iteration budgets and wall caps differ by campaign; Section 4 states their clocks and limits. Code availability and the companion register distinguish released source/evidence from unreleased runners and external binaries.

Method	Method origin	Executed lineage	Pinned version / config	Use
LR-TA	Local-ratio ordering; project specification	Project code; no external executable authentication	Source `40209c26`; frozen options	03–06, 12/13, 28
WMSF/L2	Stable-FAS context [7]	Project L2 code; external authors' binary not used	Source `40209c26`; frozen options	03–06, 12/13, 28
IPSNS	Vahidi's earlier SSRN preprint (2026)	Project code; EXP28 adds event callbacks	Baseline `40209c26`; adapter `0c8d88e2`	03–06, 10, 28
Weighted Eades	Eades et al. [14]	igraph FAS; project DAG order and rescore [23]	igraph 1.0.0; Eades option	04–06, 12/13, 28
Single-node insertion	Classical insertion neighborhood; cf. block search [37]	Project single-node search from LR-TA or WMSF	EXP05 `f13cbe50`; move/time caps	05, 12/13, 28
RASstar/FASDC arm	Xiong et al. [47]	Original-author C++ source, one FASDC branch	`8340e163`; $r = 1/3$, `-p`, original weights	10, 12/13, 28
RASstar restarts	Xiong branch [47]; study protocol	Project wrapper repeats original binary	Same source; EXP28 wall cap	28
HiGHS MILP	Linear-ordering MILP [19]	Project model via SciPy/HiGHS [39]	SciPy 1.17.1; 120-s cap	06
Gurobi SCC MILP	Linear-ordering MILP [19]	Project SCC/lazy cuts; Gurobi solver [20]	Gurobi 12.0.3; 420-s cap, four threads	07
igraph IP check	igraph exact-IP option [23]	Third-party solver as independent cross-check	igraph 1.0.0; 12 separate solves	06

and seven comparator tasks per graph, including separate single-run and restart arms of the authenticated original-author RASstar configuration [47]. The controller starts its monotonic clock before worker startup, imports, graph parsing, initialization, scoring, and checkpoint I/O. Long task caps are 120 s for synthetic and 300 s for empirical graphs. Evaluation horizons are 30 and 120 s for synthetic graphs, and 120 and 300 s for empirical graphs, extracted from the same trajectory. A budget objective is the best independently validated complete order whose order and event files were published by that horizon; absence remains missing. The direct paired comparison uses IPSNS seed 1 and the same graph/clock horizon; seed 2 is a repeated-run check. A task that reaches its cap with a valid incumbent remains a TIMEOUT. The earlier fixed-iteration and pilot records retain their original classifications. Tasks ran on several Wulver compute nodes; the manifest CPU model describes the login node, not verified per-task compute hardware. Equal horizons therefore support within-cohort quality comparisons, not cross-node speed-superiority claims.

EXP28's roster was frozen before its executions. For each generated family it takes replicate 02 at EXP06 $(n, k) = (28, 4)$ with uniform weights, and EXP04 $(120, 3)$ and $(480, 8)$, where k is target mean out-degree. The EXP25 candidate design used replicate 02 to avoid the earlier EXP11 views and span size/degree regimes. Its empirical subset uses CollegeMsg's first-90-day count view, Email-Eu-core's largest SCC, the 300-vertex U.S. 2022 EXIOBASE matrix, and U.S. OpenFlights routes. OpenFlights was a declared feasible-parent fallback for LODES initialization. This is a purposive subset of available validated inputs. The frozen protocol states that method outcomes were not used, but earlier quality and feasibility records were available; independent outcome blinding and the full rationale for each empirical-view inclusion or omission cannot be established from the preserved records.

4.5 Verification and reproducibility

Validators implemented separately from the algorithm code verify graph identity and declared vertices, require complete saved permutations, recompute the loop-free objective in Eq. 1, and check acyclicity after backward-arc removal. Independence here refers to implementation within the project, not external peer review. Cross-campaign checks reconcile weighted-graph and topology identities, historical selections, seed aggregation, comparison signs, and status denominators. Solver

orders, statuses, logs, and bounds are checked together to distinguish numerical certificates from unresolved gaps.

Original outcomes and derived recoveries remain separate. The Email-Eu-core recovery restored verified isolated vertices to incomplete orders and executed the dependency-skipped insertion tasks; the original failures and skips remain in the reliability inventory (Section 9). EXP28 retains one original INVALID strict-cap checkpoint and its declared replacement, a valid TIMEOUT with a predeadline incumbent. Its effective comparison selects that replacement only for the affected task and preserves the other 116 original task records. Budget eligibility requires a validated order and event published by the horizon; recorded filesystem timestamps are publication evidence, not cryptographic time attestations.

The companion evidence ledger records immutable sources, selection manifests, file-level checks, and validation receipts. These support reproduction of saved observations while preserving missing outcomes, recovery lineage, and numerical-certification limits.

4.6 Statistical analysis plan

The cross-campaign protocol fixed the new aggregation rules after existing campaign reports were available; it was not a blinded preregistration. RQ1 compares methods on the same validated weighted graph and states eligible and complete denominators. For the fresh 200-iteration IPSNS comparison, the graph-level outcome is the median of three original seeds relative to the better validated LR-TA/WMSF initializer; a graph enters this complete-case summary only if all three seeds and both initializers have valid objectives. Strict wins and ties use integer costs before rounding. The signed $(B_{200} - B_{\text{init}})/T$ is distinct from $(B_{\text{init}} - B_{200})/B_{\text{init}}$, which is undefined for a zero initializer objective. Generated topology is the principal resampling unit; EXP06 weights and stochastic seeds are nested within it. EXP04 and EXP06 intervals use 2,000 fixed-seed topology bootstrap replicates within family–size–design cells and describe only those synthetic designs. EXP07 shares EXP03/04 graphs and adds no independent units. EXP05's four source parents are too few for a precise parent-population interval, so parent-specific descriptive contrasts take priority over a narrow bootstrap from 38 views. RQ2 uses within-family features and an EXP04 size/target-degree-cell-centered rank sensitivity check; the 72 inspected controlled feature/outcome combinations are exploratory without multiplicity adjustment. No predictive model met the topology-grouped validation and simple-baseline requirements. RQ3 stratifies solver-certified point gaps from bound intervals and leaves zero-denominator relative gaps undefined. RQ4 retains original outcome categories and completed-only process times; a censored time-to-completion analysis requires additional assumptions. Every interval names its independent unit. EXP10 uses selected topology-level two-seed medians and reports real-view timeouts separately. EXP12 reports within-collection graph counts without assigning independent-source-parent status to each Ali file; EXP13 reports each empirical parent separately. EXP14 compares saved orders as objective-reuse evidence, not a controlled method competition. EXP16 found the available parent count, missing outcomes, and budget compatibility insufficient for leakage-safe learned algorithm selection; no predictive model is estimated here.

EXP28's direct IPSNS–RASstar comparison uses one prespecified seed per method arm and the same graph and wall-clock horizon. Its nine selected generated topologies, rather than their tasks, are the synthetic units; the displayed bootstrap interval resamples these nine whole topologies 10,000 times and describes this selected set only. Each empirical graph represents one selected source parent, so four parent-specific signs take priority over an empirical confidence interval. The two horizons on a graph are dependent views of one execution. A TIMEOUT with a verified incumbent contributes an available objective while remaining incomplete; no objective is assigned before its first saved order. The per-task completion rate, first-order availability, and best-order quality are distinct outcomes.

4.7 Public saved-output replay

The versioned public releases support saved-output replay of all ten printed numerical tables and Figure 2. Independent anonymous replays recomputed the displayed values from graph,

order, status, and selection evidence, using exact upstream transformations for third-party graphs that are not redistributed. This reproduces recorded observations rather than rerunning the optimizers; identical stochastic outcomes or runtimes are not promised. The Data and Code availability statements identify the immutable releases and reproduction instructions.

The public coverage register defines 17 claim groups as evidence-tracking categories. Eleven are publicly replayable: the ten numerical-table groups and the Wikimedia saved-order group. RQ1, RQ3, RQ4, and EXP10 remain internally verified; RQ2 is scientifically unresolved; HIST has public metadata only. The 11/17 fraction is an audit-coverage indicator, not a statistical measure of paper quality or validation of every broader claim. Exact reconstruction of the original Hugging Face inputs and pre-May-2026 Wikimedia source snapshots remains unresolved; present-day Ali inputs verify EXP12, not the older Hugging Face runs.

Figure 1 summarizes how graph identity, source parentage, saved solutions, and outcome status enter a comparison. The metadata preserve these relationships even where source-specific terms restrict redistribution.

4.8 Declaration of generative AI and AI-assisted technologies

The author used ChatGPT and Codex (OpenAI) to assist with research workflows, programming, manuscript preparation, and technical review. Gemini (Google) and Perplexity AI were used to search for relevant scientific literature and references, provide simulated critical reviews of the manuscript, and suggest improvements to its clarity, organization, and writing.

AI-generated suggestions and assessments were used as research and editorial assistance, not as substitutes for independent scientific evidence or formal peer review. The author retains full responsibility for the manuscript's scientific content, methodology, results, references, and conclusions.

5 Benchmark characterization

5.1 Graph scale and density

Table 3 reconciles the independently hash-checked campaign inputs and original tasks. The 416 distinct weighted graph hashes comprise 279 generated topologies plus 38 real-world view topologies, giving 317 distinct topology hashes; this count describes file identity, not 317 independent draws from one population. EXP06 contributes two weighted realizations per topology. EXP07 reuses 45 generated inputs and is excluded from the additive totals. Table 4 distinguishes additional contract-scoped graphs and outcome rows from the unchanged historical Hugging Face release. Table 5 retains the more detailed independently audited EXP05 parent inventory. Here $n = |V|$, $m = |A|$, and directed density is $d = m/[n(n-1)]$ for $n > 1$, after the documented treatment of self-loops. The ranges describe views within a parent, not uncertainty intervals. Dense matrix-derived views occupy a different density regime from sparse networks.

5.2 Structural properties

The cross-campaign audit independently recomputed features on all 416 weighted inputs from verified graph bytes: largest-SCC fraction, nontrivial-SCC count and sizes, total-degree coefficient of variation (CV), reciprocity as the fraction of arcs with a reverse arc, and positive-arc-weight CV. CV is undefined when its mean is zero. SCCs retain isolated vertices in the declared vertex set [40]. EXP06's paired weights share unweighted structural properties and do not supply independent structural observations.

6 Cross-family heuristic performance

6.1 Overall comparative solution quality

RQ1 uses paired differences on graphs with two validated complete orders under a stated computational protocol. Pairing weighted inputs controls graph identity, not work: the fixed-iteration and one-shot comparisons below are descriptive at their recorded budgets; only the EXP28 cohort matches end-to-end horizons. For comparator a

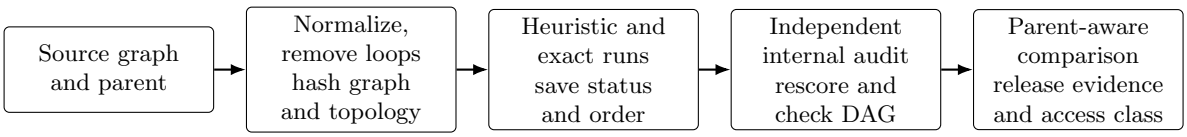


Fig. 1 Evidence workflow for one graph view. Graph and topology hashes distinguish identical weighted inputs from paired reweightings; source-parent IDs identify dependent views. Saved complete orders, original run statuses, and numerical certificate classes are checked before comparison. Public metadata preserve the joins; released orders and released or exactly reconstructed graph bytes support saved-output replay of all ten numerical tables and Figure 2.

Table 3 Hash-reconciled cross-campaign inventory. Original task records include methods, seeds, and budgets, and are not independent graph observations. The 1,398 EXP06 tasks comprise 1,386 principal heuristic/HiGHS tasks and 12 separate igraph IP cross-checks. EXP07 contains 45 selected exact cases whose graph hashes exactly match nine EXP03 and 36 EXP04 inputs; its row is not added to the EXP03–06 totals. EXP06 shares neither graph nor topology hashes with EXP07. Source: independent cross-campaign audit, `canonical_graphs.csv`, `canonical_runs.csv`, and `exp07_hash_overlap.csv`, release `08a2a48`.

Campaign	Weighted inputs	Topology hashes	Original tasks	Dependence
EXP03	45	45	900	One weight per topology
EXP04	135	135	2,025	Five replicates per design cell
EXP05	38	38	760	Four real-world parents
EXP06	198	99	1,398	Two weights per topology
EXP03–06 total	416	317	5,083	Repeated tasks on distinct inputs
EXP07, reused	45	45	—	9 EXP03 + 36 EXP04 matches

against LR-TA on graph g, define $\Delta_{a,L}(g) = [B_g(a) - B_g(L)]/T_g$, where T_g is total retained arc weight. Negative values favor a; a strict win, tie, or loss is determined by the unrounded integer objective, not the displayed decimal. This normalized difference is a fraction of total graph weight, *not* a percentage improvement relative to LR-TA's objective. For iterative refinement, Table 6 instead uses the better recorded LR-TA/WMSF initializer as its same-graph reference. The audited EXP05 original inventory contains 722 planned comparator-run pairs, 601 complete and 121 marked not comparable because at least one original result was incomplete. These are run pairs, including repeated IPSNS seeds, rather than 601 independent graphs. A complete-case summary describes observed orders only; it cannot reconstruct the ranking had timed-out runs completed.

6.2 Family-level results

Table 6 shows the independently checked fresh 200-iteration EXP04 comparison. Its zero-loss counts follow incumbent protection, rather than an independent empirical advantage. The informative quantities are strict-improvement frequency and magnitude, ties, and additional runtime cost. Each family has 45 complete topology-level three-seed sets. The median signed reduction is largest for Erdős–Rényi and smallest for scale-free inputs under this protocol. For scale-free graphs the median $\Delta/T = -0.000415$ means 0.0415 percentage points of total graph weight, whereas the separate median improvement relative to the initializer's backward cost is 0.816%. Erdős–Rényi has a 0.662-percentage-point total-weight reduction and a 4.48% baseline-relative reduction. The paired process-wall ratios quantify the additional cost of these gains; Section 9 examines their budget dependence. Across nine matched EXP04 size/target-degree cells, the scale-free minus Erdős–Rényi median signed-effect difference is +0.006869 and is positive in all nine cells. This conditional generator contrast is descriptive, not a causal family effect or a universal heuristic ranking. EXP03's one-topology-per-cell cohort has the same qualitative median ordering of refinement effects, but lacks within-cell replication.

Under the EXP04 uniform-weight generator, the evaluated weighted igraph/Eades configuration loses all 135 same-input pairs against LR-TA, with family median signed differences between

Table 4 Audited extension inventory, separate from the unchanged 7,371-row/14-configuration public historical release. Graph rows are unique only within their stated normalization contracts; the total of 442 is not a global distinct-graph count. Outcomes include repeated methods and seeds and preserve incomplete statuses. EXP07's 45 numerical-reference rows reuse EXP03/04 graphs and are excluded from outcomes. Source: EXP16 identity summary and frozen private EXP18 candidate at commit `8fae3ea`; its path-free metadata projection was subsequently released publicly at the revision stated in Data availability.

Cohort	Graph rows	Outcome rows	Independent-unit qualification
EXP03–06 common contract	416	5,083	317 topologies; four EXP05 empirical parents
EXP12 Ali collection	20	140	Source-parent independence unresolved
EXP13 new empirical sources	3	21	Three distinct source parents
EXP14 Wikimedia history	3	10	One source parent; three dependent views
Contract-scoped total	442	5,254	No verified global graph deduplication
EXP07 references, reused	45	—	20 numerical optima; 25 bounded cases

Table 5 Audited EXP05 real-world benchmark inventory. Ranges are minima and maxima across graph views within each source parent; m counts retained directed arcs and $d = m/[n(n-1)]$. The 38 views arise from four sources, not 38 independent source datasets. Source: EXP05 independent audit, benchmark inventory, revision `559f841`.

Source parent	Views	n	m	d	Arc-weight convention
CollegeMsg	4	1,086–1,899	8,111–20,296	0.00563–0.01137	Message counts
Email-Eu-core	2	803–1,005	24,138–24,929	0.02471–0.03748	Unit
Wiki-Vote	2	1,300–7,115	39,456–103,689	0.00205–0.02336	Unit
EXIOBASE matrices	30	10–300	43–44,356	0.32696–0.50000	Published integer weights

+0.0202 and +0.0496 of total arc weight. EXP06 isolates a different design factor: two weight regimes on each of 99 fixed generated topologies. Table 7 shows that this same weighted Eades implementation has a higher solver-certified relative gap under capped-lognormal weights on most paired topologies in each family. These are within-topology differences in $(H - \mathrm{OPT})/\mathrm{OPT}$, not raw-objective differences or percentages of total graph weight. Other methods have smaller or mixed paired effects in the audit table; the result is specific to this implementation and synthetic reweighting, not a universal property of heavy-tailed weights or all feedback-arc heuristics.

6.3 Real-world results

Table 8 gives descriptive, original complete-case comparisons by source parent for four deterministic comparators. The 120-second task cap applies to all EXP05 tasks, but algorithmic stopping and initialization costs differ; these are quality comparisons, not equal-compute performance rankings. CollegeMsg's four views show four strict improvements for insertion from LR-TA, with median $\Delta_{a,L} = -0.005467$ (0.5467 percentage points of total arc weight). On the 30 related EXIOBASE views, the same comparison gives 19 wins and 11 ties, with median −0.000578 (0.0578 percentage points of total arc weight). The latter magnitude is small on this normalization, and neither count estimates a population effect over independent datasets. Weighted Eades loses to LR-TA on all four completed CollegeMsg views; WMSF/L2 loses on the one comparable Email-Eu-core view. Email-Eu-core has only one of two original LR-TA pairs available because the full-graph LR-TA ordering originally failed validation. The separate Email-Eu-core recoveries are discussed with reliability in Section 9; they are not inserted into this original complete-case table.

6.4 Original-author comparator under unequal work

EXP10 supplies a matched external-implementation check rather than another independent graph sample. On 18 selected EXP04 synthetic topologies, two original-Xiong RASstar and two fresh IPSNS orders per graph passed

Table 6 Fresh 200-iteration IPSNS refinement on EXP04. C/E is complete/eligible generated topologies; W/T/L compares the median objective of three original IPSNS seeds with the better validated LR-TA/WMSF initializer on the same graph. $\Delta = (B_{200} - B_{\text{init}})/T$ is signed and normalized by total arc weight; negative favors refinement. The 95% bootstrap interval for median Δ resamples topologies within nine size/target-degree cells (2,000 replicates) and describes this synthetic design. Baseline improvement is the separate median $(B_{\text{init}} - B_{200})/B_{\text{init}}$, expressed as a percentage. The wall ratio is the median of per-graph recorded process-wall ratios, not an equal-budget speed ratio. Source: cross-campaign audit `quality_summary.csv` and `refinement_summary.csv`, release `08a2a48`.

Family	C/E	W/T/L	Median Δ	Bootstrap interval	Baseline gain	Wall ratio
Erdős–Rényi	45/45	39/6/0	−0.006616	$[-0.008054, -0.004559]$	4.48%	9.63
Community	45/45	35/10/0	−0.003955	$[-0.005957, -0.002380]$	2.91%	9.82
Scale-free	45/45	27/18/0	−0.000415	$[-0.001014, 0]$	0.816%	3.19

Table 7 EXP06 fixed-topology sensitivity of the evaluated weighted Eades implementation. The difference is capped-lognormal minus uniform solver-certified relative regret for the same topology; positive means a larger gap under capped-lognormal weights. Higher/equal/lower counts use exact paired differences. Intervals are 2,000-replicate bootstrap intervals for the median, resampling topologies within size/degree or planted-order design cells. The small cells and synthetic generator restrict generalization. Source: cross-campaign audit `exp06_weight_pairs.csv` and `exp06_weight_summary.csv`, release `08a2a48`.

Family	Paired topologies	Higher/equal/lower	Median gap difference	Bootstrap interval
Erdős–Rényi	27	22/0/5	+0.230	$[0.120, 0.324]$
Scale-free	27	20/2/5	+0.146	$[0.119, 0.356]$
Community	27	21/1/5	+0.112	$[0.072, 0.296]$
Planted order	18	13/0/5	+0.494	$[0.327, 0.592]$

independent permutation, objective, and retained-DAG validation. IPSNS had lower backward weight on all 18 topology-level two-seed medians and all 36 seed pairs. The median graph-level $(B_{\text{Xiong}} - B_{\text{IPSNS}})/T$ was +0.025215, or 2.5215 percentage points of total weight; the separate median excess relative to the IPSNS objective was 0.183790. On the 36 comparable synthetic tasks, median recorded end-to-end process wall time was 0.356 s for RASstar and 7.282 s for IPSNS. These values describe two unequal-work implementations, not an equal-compute speed/quality frontier. On the three reused EXP05 full views, all six RASstar tasks produced valid orders while all six 200-iteration IPSNS tasks reached the 120 s cap; the absent IPSNS objectives preclude a matched quality statement there. A separately audited, post hoc cross-campaign join found RASstar lower than the best eligible archived EXP05 objective on each of those three graph bytes, but different budgets and selection rules prevent a controlled ranking.

6.5 Matched wall-clock comparison

EXP28 addresses the unequal-time and missing-incumbent limitations with complete-order checkpoints under one end-to-end clock. Table 9 compares IPSNS seed 1 with the authenticated RASstar single-run and restart arms on exactly matched graph bytes and prespecified horizons. On nine selected generated topologies, IPSNS beats the single-run arm on seven, ties on one, and loses on one at both 30 and 120 s; its median total-weight benefit is 1.096 percentage points. The mean benefit is 1.255 points, with a descriptive whole-topology bootstrap interval of $[0.489, 2.081]$ points. The signs differ by family: three wins on Erdős–Rényi, two wins and one loss on community, and two wins and one tie on scale-free inputs. The selected three topologies per family do not estimate a family-wide win rate.

Across four selected empirical source parents at 120 and 300 s, IPSNS wins only on the EXIOBASE view (1.443 percentage points of total weight) and loses on CollegeMsg (0.144), Email-Eu-core (2.167), and OpenFlights (0.285). Negative values in Table 9 favor RASstar. Its restart

Table 8 EXP05 original complete-case quality against LR-TA, grouped by source parent. A comparator win means a strictly smaller backward-arc weight. C/E is complete/eligible graph views; W/T/L are strict comparator wins/ties/losses. Median Δ is median$[(B_{\text{comparator}} - B_{\text{LRTA}})/T]$ over the C completed views; negative favors the comparator. Missing original pairs remain in E. IPSNS repeated seeds are omitted. Values are descriptive across dependent views under unequal algorithmic effort, with no confidence interval. Source: EXP05 independent audit, `performance_by_parent.csv`, revision `559f841`.

Source parent	Comparator	C/E	W/T/L	Median Δ
CollegeMsg	WMSF/L2	4/4	3/0/1	−0.000169
CollegeMsg	Weighted Eades	4/4	0/0/4	+0.018635
CollegeMsg	Insertion from LR-TA	4/4	4/0/0	−0.005467
CollegeMsg	Insertion from WMSF	4/4	4/0/0	−0.006156
Email-Eu-core	WMSF/L2	1/2	0/0/1	+0.042174
Email-Eu-core	Weighted Eades	1/2	1/0/0	−0.006546
Email-Eu-core	Insertion from LR-TA	1/2	1/0/0	−0.021584
Email-Eu-core	Insertion from WMSF	1/2	1/0/0	−0.020300
Wiki-Vote	WMSF/L2	2/2	0/0/2	+0.013473
Wiki-Vote	Weighted Eades	2/2	0/0/2	+0.011787
Wiki-Vote	Insertion from LR-TA	2/2	2/0/0	−0.007067
Wiki-Vote	Insertion from WMSF	2/2	1/0/1	+0.001079
EXIOBASE	WMSF/L2	30/30	6/9/15	+0.000032
EXIOBASE	Weighted Eades	30/30	0/2/28	+0.029486
EXIOBASE	Insertion from LR-TA	30/30	19/11/0	−0.000578
EXIOBASE	Insertion from WMSF	30/30	19/10/1	−0.000400

arm changes two empirical objectives between horizons but does not change the one-win/three-loss sign count. All thirteen seed-1 single-run pairs are observed at both horizons; no missing objective is silently scored as a loss. These empirical views differ from EXP10's three larger full-graph inputs, so the new availability evidence does not retroactively create EXP10 same-input comparisons.

Figure 2 shows each graph's single-run paired effect at both prespecified horizons. The flat segments are observed stability of the best validated orders, not evidence that either method could never improve with further work. The graph-level view makes the community loss, scale-free tie, and parent-specific reversals visible alongside the aggregate table.

The same audited matched-clock cohort also includes project single-node insertion. At 300 s, insertion from either LR-TA or WMSF has lower backward weight than IPSNS seed 1 on all four empirical parents; at 120 s, each wins on three and loses on one. On the nine synthetic topologies, IPSNS versus LR-TA insertion has three wins, two ties, and four losses at both horizons; versus WMSF insertion it has three of each. Both median synthetic paired differences are zero. Thus the selected synthetic advantage over RASstar does not extend to every evaluated comparator.

6.6 Historical benchmarks and additional empirical parents

EXP12 evaluates 20 distinct present-day Ali graph hashes with 140/140 independently valid task outcomes. The source bytes of all 20 also occur in an earlier IPSNS experiment ledger; their contribution here is separately audited comparator outcomes and graph-identity linkage, not new graph inputs. Eighteen of 20 graphs have at least two best-observed method families, and four have zero objective spread across the tested families. LR-TA and RASstar each attain the best-observed objective on 16/20 graphs, counting ties; neither count means a certified optimum. Both RASstar seed objectives agree on all 20 graphs, although only three pairs have identical order hashes. The historical Hugging Face name and numeric crosswalk is contextual, not a byte-level replication of the original rows. EXP13 adds three genuinely separate

Table 9 Audited EXP28 IPSNS seed 1 versus original-author RASstar seed 1 under common end-to-end wall-clock horizons. W/T/L counts strict IPSNS objective wins/ties/losses on the same weighted graph. Median gain is $100(B_{\mathrm{RASstar}} - B_{\mathrm{IPSNS}})/T$ in percentage points of active total arc weight; positive favors IPSNS. A slash-separated budget row combines horizons only when every graph-level paired objective is identical at both. The nine synthetic topologies are selected reused inputs; each of the four empirical rows represents one source parent. TIMEOUT tasks with predeadline validated orders contribute objectives but remain TIMEOUT. Source: sealed EXP28 recovery audit and independent integration recalculation.

Cohort	Horizon (s)	RASstar arm	Pairs	IPSNS W/T/L	Median gain (pp)
Selected synthetic	30/120	Single	9	7/1/1	+1.096
Selected synthetic	30/120	Restarts	9	7/1/1	+1.096
Empirical parents	120/300	Single	4	1/0/3	−0.215
Empirical parents	120	Restarts	4	1/0/3	−0.308
Empirical parents	300	Restarts	4	1/0/3	−0.311

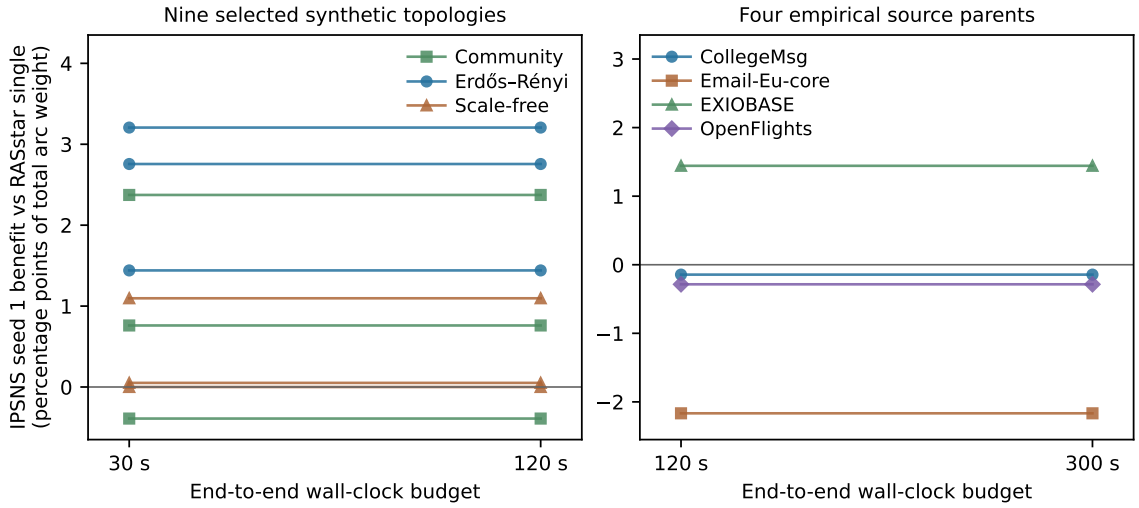


Fig. 2 Matched-budget IPSNS seed 1 benefit over authenticated RASstar single-run seed 1, in percentage points of each active graph's total arc weight. Each segment is one reused weighted graph observed at two horizons of the same run; positive values favor IPSNS. Colors and marker shapes distinguish generated families (left) and empirical source parents (right). Valid predeadline incumbents are used even when their tasks end as TIMEOUT; all displayed pairs are available at both horizons. The plot is descriptive for nine selected topologies and four selected parents, not a population performance curve. Source and deterministic generation are pinned in the EXP28 integration evidence ledger.

source parents and 19/21 valid task outcomes; its two IRS insertion tasks timed out without objectives. Table 10 shows the different best-observed winners. Across the matched EXP12/13 graphs, LR-TA has 11 strict wins, two losses, and ten ties against WMSF/L2; RASstar seed 1 against LR-TA has four wins, four losses, and 15 ties. These 23 graph-level descriptions are nested in a heterogeneous Ali collection and three empirical parents, not 23 independent source datasets. Completion, method selection, and wall clocks must be reported by cohort; no pooled runtime or population significance test follows from them.

6.7 Sensitivity and consistency

Repeated IPSNS seeds and fresh-run budgets describe sensitivity within graphs. EXP08's small selected cohort shows that search paths can differ even when final objectives agree. In EXP28, seed 2 reproduces seed 1's best objective on all 13 graphs. Such agreement describes the available results on these inputs; it does not establish general robustness or resolve timed-out objectives. Section 9 examines budget sensitivity and completeness separately.

Table 10 EXP12/13 audited method outcomes. The Ali row summarizes one collection with 20 within-collection graph identities; the other rows are one graph per independent empirical parent. PASS/assigned counts tasks, including two RASstar seeds per graph. Best-observed active-graph costs are feasible values, not certified optima and not comparable across rows with different weight semantics. Original source-loop constants are excluded from active optimization: LODES 18,628; IRS 124,358,925. Source: EXP12 `summary.json`, `graph_comparisons.csv`; EXP13 `statistical_summary.json`, `graph_method_comparison.csv`, revision of PR #25/#26.

Cohort / source parent	Graphs	PASS/assigned	Best-observed result	Interpretation
Ali collection	20	140/140	18/20 graphs: best-family ties	Four graphs: zero spread
OpenFlights	1	7/7	5,099 (RASstar seed 1)	Seven feasible orders
Census LODES	1	7/7	65,397 (WMSF insertion)	Loop constant 18,628
IRS SOI	1	5/7	2,770,227 (LR-TA)	Two insertion timeouts

7 Exact optimization and optimality gaps

7.1 Exact-solver methodology

EXP06 uses SciPy/HiGHS with pairwise precedence variables and transitivity constraints on 198 weighted instances, representing 99 topologies under two weight regimes. The cross-campaign audit independently reconciled all 198 solver statuses, feasible objectives, bounds, and graph identities; 12 separate igraph integer-programming checks agreed with the corresponding optima. EXP07 uses Gurobi 12.0.3 on 45 selected EXP03/04 graphs with SCC decomposition and lazy triangle constraints. Its audit validated 270 historical heuristic records and formed 180 graph–algorithm comparisons using the historical best record per method and graph. It rechecked solver statuses and the per-SCC integer minimization-bound convention $\max\{0, \lceil \text{ObjBound} - 10^{-4} \rceil\}$ before summing component bounds. The adjustment guards rounding at floating-point thresholds but is not an exact-arithmetic proof. EXP06 and EXP07 share neither graph nor topology hashes, and their inputs, limits, and solver protocols do not support a paired solver comparison.

7.2 Certified-optimal cases

All 198 EXP06 weighted instances are solver-certified numerical optima on small graphs with $n \in \{12, 20, 28\}$; none has a zero optimum. At the weighted-instance level, 85/198 LR-TA, 77/198 WMSF/L2, 17/198 weighted Eades, and 115/198 IPSNS median-of-three objectives equal the certified optimum. Their respective median relative regrets are 0.0170, 0.0196, 0.2937, and 0. The repeated IPSNS seeds and two weights per topology are not additional independent topology units. These small-instance rates do not estimate optimality frequencies for the larger campaigns.

Separately, 20 of 45 EXP07 selected cases are solver-certified under audited Gurobi numerical conventions. The median relative regret of historical heuristic records on these 20 cases is 0.0472 for IPSNS, 0.0880 for LR-TA, 0.1483 for WMSF/L2, and 0.4393 for weighted Eades (Table 11). Historical best-record selection and graph sizes differ from EXP06's fixed 200-iteration three-seed summary; these medians are not pooled or compared as a matched cohort. Certification rests on solver status, complete order, bound-log agreement, and the stated convention, not Slurm success or equality with a feasible incumbent.

7.3 Uncertified instances

The other 25 EXP07 cases have feasible orders and positive validated lower bounds, but each reached the solver time limit with an open gap. None is an unsolved or failed graph. Equation 2 yields case-level regret intervals; Table 11 reports the medians of their lower and upper relative endpoints separately. For IPSNS these medians are 0 and 1.874, illustrating how little a feasible incumbent alone can establish when lower bounds are loose. These two endpoint medians are not an observed median exact gap or a confidence interval. The original aggregate's `timeouts: 0` counted only top-level worker timeouts; the audit separately documented 25 capped solver terminations.

Table 11 EXP07 heuristic regret in separate numerical-certification classes. Each method has 20 selected graph inputs with solver-certified optima and 25 with a feasible incumbent and validated lower bound but no optimum certificate. Certified entries are medians of $(H - \mathrm{OPT})/\mathrm{OPT}$ across the 20 positive-optimum cases. Open entries are *separate medians* of per-case lower and upper relative interval endpoints from Eq. 2; they are neither observed exact gaps nor confidence limits. Historical best heuristic records are used, not EXP06's fixed 200-iteration protocol. Source: EXP07 independent audit and cross-campaign `exp07_gap_classes.csv`, `exp07_gap_summary.csv`, release `08a2a48`.

Certificate class	Method	Graphs	Certified median	Lower median	Upper median
Certified	LR-TA	20	0.0880	—	—
Certified	WMSF/L2	20	0.1483	—	—
Certified	Weighted Eades	20	0.4393	—	—
Certified	IPSNS	20	0.0472	—	—
Uncertified	LR-TA	25	—	0.0385	1.9695
Uncertified	WMSF/L2	25	—	0.0813	2.1008
Uncertified	Weighted Eades	25	—	0.3992	2.7551
Uncertified	IPSNS	25	—	0	1.8735

8 Structural analysis of heuristic difficulty

8.1 Structural feature definitions

We distinguish a graph feature, a heuristic's observed quality loss, and an exact solver's effort. Features computed before algorithm outcomes include n, m, density, largest-SCC fraction $|C_{\max}|/n$, nontrivial-SCC count, degree CV, reciprocity, and weight CV. Heuristic difficulty is a validated paired loss or certified/bound-based regret under fixed method and budget. Solver effort is optimization time, logged search work, or certification at the cap. These outcomes are not synonyms.

8.2 Feature–performance relationships

The audited EXP05 ranges establish variation in size and density, and the full Email-Eu-core graph has 19 isolated vertices whose omission explained its original incomplete orders. These observations describe inputs and a validation failure; they do not establish that either characteristic predicts heuristic objective loss. The independent cross-campaign reader recomputed structural features from every verified graph input and joined them to complete same-input quality contrasts. As one exploratory EXP04 example, within its 45 scale-free graphs the unadjusted rank association between largest-SCC fraction and weighted Eades' signed disadvantage to LR-TA is +0.830. A size/target-degree-cell-centered rank check gives +0.340 over nine variable cells. The attenuation warns against reading the unadjusted coefficient as a stable structural explanation. It neither isolates a mechanism nor predicts behavior on new graph families.

8.3 Family-controlled analysis

Size, density, family, and SCC coverage can move together under the generator. The audit examined 72 EXP04 family–method–feature combinations after centering ranks within size/target-degree cells; the example above was selected from that set, without multiplicity adjustment. Each cell contains only five generated topologies, and several features are correlated or constant within cells. Thus these associations are exploratory, with no confirmatory interval or predictive model. EXP07's size/SCC–solver-effort relationship is a different, cap-affected outcome and must not be substituted for heuristic quality difficulty. Four EXP05 source parents are inadequate for adjusted real-world population inference. A future predictive claim would require topology-grouped or source-parent-held-out evaluation against a family/size-only baseline; the present data do not establish it.

9 Runtime, reliability, and trade-offs

9.1 Runtime distributions

Runtime comparisons retain the clock definitions and computational protocols in Section 4. Table 6 compares full-process wall time with the selected initializer, including overhead beyond refinement itself. A common task cap equalizes available time rather than work. Raw times are not pooled across hardware or protocols, and EXP07's 25 capped solver cases are excluded from completed-time-to-certification summaries.

Pairing the original EXP04 records by graph hash and seed gives 405 complete groups with identical objectives and vertex permutations at 10, 50, 200 and 400 iterations; all 1,620 runs are PASS. EXP03 likewise has 135 such groups across 10, 25, 50, 100, 200 and 400 iterations (810 PASS runs). These are fresh runs, not checkpoint trajectories, and equality was checked before aggregate rounding. EXP04's completed process-wall median rises from 1.441 s at 10 iterations to 12.285 s at 400; individual timings need not increase monotonically. For each graph, we select the lower-objective initializer, breaking objective ties by recorded wall time, and divide the three-seed median IPSNS process wall by that initializer's wall. Taking family medians reproduces Table 6's aggregation: the 10-iteration ratios are 1.60, 1.53 and 1.27 for Erdős–Rényi, community and scale-free inputs, versus 9.63, 9.82 and 3.19 at 200 iterations. All 135 graph comparisons are complete; the deterministic analysis and manifest-bound inputs are identified in `analysis/manuscript_clarity/runtime_ratios.py` in the companion repository. The higher budget therefore increased recorded cost without changing the saved solutions in these paired groups. Ten iterations sufficed for their observed quality, but this plateau establishes neither general convergence nor an adaptive stopping rule.

9.2 Failures, censoring, and timeouts

Table 12 retains every original EXP05 outcome. All 108 TIMEOUT records belong to IPSNS and are right-censored at the 120-second task cap: their eventual completion times and objectives are unknown. By source parent, the original timeout counts are 21/80 CollegeMsg tasks, 12/40 Email-Eu-core, 21/40 Wiki-Vote, and 54/600 EXIOBASE; these denominators are tasks on related views, not independent datasets. The 11 FAIL records are incomplete Email-Eu-core full-graph orderings, not slow completions; the two insertion tasks were skipped because their initializer orders failed. Thus IPSNS completed 453 of 570 original runs, whereas each one-shot LR-TA and WMSF method completed 37 of 38; these counts describe different task definitions and budgets. At the fresh 200-iteration budget, all three original IPSNS seeds and both initializers yield a complete graph-level comparison for only 18 of 38 EXP05 views, all from the EXIOBASE parent. Two of these 18 strictly improve on the better initializer and 16 tie; their median signed total-weight effect is zero. The remaining 20 views do not have complete three-seed evidence, so this complete-case result cannot describe all four parents. Survival-style analysis would need a common task definition and defensible censoring assumptions; otherwise the original status distribution is the more faithful reliability result.

The separate recovery campaign appended exactly 19 independently verified isolated vertices to each of the 11 affected orders without changing the non-isolated order or weighted objective. It then executed only the two skipped insertion tasks from the validated initial orders. All 13 derived endpoints passed independent complete-order, objective, DAG, and provenance checks, but their original FAIL and SKIP labels remain. On the Email-Eu-core full graph, insertion from LR-TA reduced backward weight from 10,095 to 9,607 (488 units), and insertion from WMSF/L2 reduced it from 11,134 to 9,699 (1,435 units). These are two derived runs on one existing graph, not new independent instances or original-campaign successes.

9.3 Solution-quality versus runtime trade-offs

The derived LR-TA insertion used 98.570 seconds of refinement and 102.014 seconds of subprocess wall time; adding its earlier initializer task gives a 108.529-second composite. The corresponding WMSF insertion used 110.257 seconds of refinement and 112.590 seconds of subprocess wall

Table 12 Complete original EXP05 reliability inventory, before recovery. PASS means an independently validated complete order; FAIL means an invalid original ordering; TIMEOUT is right-censored at the 120-second task cap; SKIP is an original dependency skip. IPSNS has 15 fresh seed–budget runs per graph; each other method has one original task per graph. Derived results are excluded from all counts. Source: EXP05 independent audit, `original_inventory.csv` and `reliability_by_algorithm.csv`, revision `559f841`.

Method	Total	PASS	FAIL	TIMEOUT	SKIP
LR-TA	38	37	1	0	0
WMSF/L2	38	37	1	0	0
Weighted Eades	38	38	0	0	0
Insertion from LR-TA	38	37	0	0	1
Insertion from WMSF	38	37	0	0	1
IPSNS	570	453	9	108	0
All original tasks	760	639	11	108	2

time, with a 123.935-second initialization-inclusive composite. The 110-second refinement deadline is checked between sweeps, so the latter refinement slightly exceeded it while the 120-second subprocess cap held. The WMSF run ended at that deadline. Although the LR-TA run recorded a local-optimum termination, its search trace was not independently replayed, so only its feasible endpoint and objective improvement are validated here. These composite times join separate tasks and must not be called single-run wall times. Together with EXP04's objective effects and process-wall ratios, these observations show why refinement gains must be assessed alongside their initialization and execution costs.

EXP28's effective 117-task inventory has 98 PASS and 19 TIMEOUT, each timeout retaining a validated predeadline incumbent. All 117 tasks have a complete order by their smaller prespecified horizon. Both IPSNS seeds completed on all nine synthetic graphs; on the four empirical parents, two of eight IPSNS tasks completed and six timed out with usable orders. All 13 continuous RASstar-restart tasks timed out with usable orders; each of the 13 separate single-run tasks completed. For the paired seed-1 arms, median time to first published order was 0.958 s for IPSNS versus 0.676 s for RASstar on synthetic inputs, and 11.348 versus 2.776 s on empirical inputs. These medians summarize nine selected graph pairs or four selected parents, not independent seed replicates. No IPSNS task improved its best validated objective between its smaller and larger horizon; two empirical restart tasks did improve after 120 s without changing the empirical win/loss pattern. The additional time therefore did not improve IPSNS quality on these selected inputs, although the deadline still governs completion status. Equal wall-clock horizons equalize opportunity, not algorithmic work or the number of completed runs.

9.4 Implications for algorithm selection

For the tested EXP04 generators, additional IPSNS work is most defensible when the observed family-specific reduction matters relative to its time cost; on scale-free inputs the median reduction is small and 18/45 graph-level comparisons tie. The tested ten-iteration results materially reduce that observed cost assessment, but do not predict what longer runs will attain on a new graph. A practitioner must still check graph regime, input weights, available time, and the need for a certificate. Limited source-parent diversity and incomplete or incompatible method/budget coverage restrict these observations to evaluation guidance; no algorithm-selection rule is trained or validated.

10 Discussion

10.1 Synthesis across research questions

The joined evidence changes three evaluation decisions. First, EXP04's identical saved orders across fresh budgets make its refinement-cost assessment budget-dependent: reporting only 200 iterations

obscures the lower observed cost at ten. Second, the 18 complete EXP05 comparisons at 200 iterations all come from EXIOBASE; their two strict improvements and 16 ties describe available views of one parent, not all four parents or 18 independent sources. Third, EXP28's selected synthetic advantage over RASstar does not imply an advantage over insertion: synthetic median effects against both insertion initializers are zero, and at 300 s both insertion arms beat IPSNS seed 1 on all four selected empirical parents. Comparator choice and cohort therefore change the comparative conclusion. These purposive cohorts and dependent horizons cannot support a pooled population ranking; obtaining timely incumbents in EXP28 also does not fill missing objectives in the earlier, different protocol.

Topology-paired reweighting shows that structure alone does not fix the evaluated Eades implementation's regret. Numerical certification determines whether that regret is a point value or only bounded. Retaining original failures and recording recoveries separately prevents missingness from masquerading as a method loss or a new observation. These are empirical uses of the combined resource that metric rows alone cannot support; the structural correlations remain exploratory.

10.2 Practical implications

For a new weighted graph, verify the input transformation and objective convention, then compare complete orders under an application-relevant budget. Report both total-weight savings and improvement relative to the initializer: a frequent strict win may still have a small practical effect. A cheap initializer or RASstar order may be useful when completion is paramount; refinement is useful when its measured gain justifies its cost. Test the intended weight regime rather than transferring a uniform-weight ranking. When certification matters, report feasible and lower bounds together; matching an incumbent is not an optimum certificate.

10.3 Contemporary comparator scope

The authenticated RASstar $r = 1/3$ branch supplies a contemporary original-author comparator, with single-run and restart arms distinguished. It does not represent the whole FASDC portfolio. The incomplete five-view pilot remains outside the matched-wall analysis. Ben Ali's population search, Qian's block insertion, the full Rocket-Crane pipeline, and external Cavallaro/TIGHT-CUT executables have no verified same-input outcomes here; project single-node insertion cannot stand in for those methods. MA-EDM's separate feasibility pilot does not close this comparative gap, and no performance claim is made for its 2025 successor. This scope supports the reported conditional comparisons without establishing superiority over contemporary methods generally.

10.4 Threats to validity

Construct validity depends on complete vertices, self-loop treatment, matrix orientation, and weight semantics. Source-loop constants belong only in original-source totals, not active-graph comparisons. External validity is limited by the controlled generators and selected exact and matched-wall cohorts. EXP28 selection was purposive and feasibility-conditioned; a frozen roster does not establish outcome-blind sampling. Views, paired weights, seeds, and horizons are dependent; Ali source-parent independence remains unresolved. Topology-bootstrap intervals describe the selected synthetic designs, not all directed networks, and four matched-wall empirical parents do not support a precise population interval.

The 72 controlled structural associations lack multiplicity adjustment and independent predictive validation. Solver certificates rely on numerical tolerances; open-case bounds may be loose. Unequal work and hardware records limit speed comparisons, including any cross-node interpretation of matched-wall availability times. Missing objectives remain missing, while valid timeout incumbents retain censored completion times. Complete long-budget EXP05 comparisons concentrate in EXIOBASE and cannot represent its other parents. Although all ten numerical tables and Figure 2 support public saved-output replay, six broader claim groups and exact historical-source reconstruction retain the limits described in Section 4.7. These constraints preclude a universal ranking or validated selector.

10.5 Reproducibility and data reuse

Saved-output replay makes the reported comparisons independently checkable without requiring new optimization runs. It also enables reuse beyond the original metric summaries. For example, ten saved rankings on three historical Wikimedia graphs reproduce best active objectives of 3,046,487 for Arabic December 2025, 3,206,337 for Arabic January 2026, and 64,125 for Azerbaijani December 2025. On those same graphs, the saved paper-FAS and WMSF orders differ by 5,806, 20,166, and 49 objective units, respectively, with the preferred saved method reversing on Azerbaijani. These are comparisons of saved orders with unverified historical execution clocks, not controlled new runs or independent source-parent replicates.

Reproducing an objective does not establish the identity of an unavailable upstream source. Current publisher TSVs differ from the pinned derived Wikimedia graphs, leaving historical monthly source reconstruction unresolved. The same distinction separates present-day Ali evidence from the original Hugging Face runs. The coverage classifications in Section 4.7 and the source-specific access documentation make these reuse boundaries explicit.

11 Conclusions

The audited MWFAS benchmark connects exact inputs, source relationships, saved solutions, and original outcomes to comparisons of existing methods. The findings show why these connections matter: larger iteration budgets can increase recorded cost without improving saved solutions; incomplete comparisons can concentrate in one source parent; and a method's advantage can change with graph regime, weighting, or comparator. Numerical certificates and open bounds support different levels of precision in assessing heuristic quality.

For the tested inputs, refinement is best assessed through strict gains, ties, computational cost, and completion together. The resulting evidence informs budget and comparator choices within the studied cohorts, rather than a universal method ranking. Broader comparisons require more independent parents and compatible computational protocols. The contribution is a reusable empirical resource and scoped comparison, distinct from the author's earlier algorithm and ranking work.

Statements and Declarations

Data availability. The original MWFAS Heuristic Metrics release and the separate benchmark metadata and Wikimedia evidence remain available at the linked immutable revisions. Three versioned packages support the numerical results: EXP04/06 for Tables 6 and 7; EXP28 for Table 9 and Figure 2; and EXP03/05/07/12/13 for the remaining seven numerical tables. Together with the Wikimedia saved orders, these support public saved-output replay of 10/10 numerical tables, Figure 2, and 11 of 17 tracked claim groups. The other six groups retain the classifications in Section 4.7.

The packages contain rights-cleared graph/order/status evidence, original selection and recovery lineage, statistical sources, manifests, and validation receipts. Third-party graphs that are not redistributed are obtained through executable, hash-pinned upstream retrieval and exact transformations, subject to their source-specific terms. Author-generated data and metadata are CC BY 4.0; author software is MIT. These licenses do not replace third-party terms. Original Hugging Face graph/order inputs, including the historical Ali runs, and the older Wikimedia monthly TSV snapshots remain unavailable for exact reconstruction. No extension DOI is assigned.

Code availability. The project implementations of LR-TA, WMSF/L2, and baseline IPSNS are publicly pinned. Separate immutable verifier releases provide complete instructions for Tables 6/7, Table 9/Figure 2, and the remaining tables. Their documented clean-environment commands use the corresponding linked dataset revisions and fresh directories to download artifacts anonymously and recompute the reported statistics. The Python 3.12 standard-library entry points are `anonymous_replay.py`, `anonymous_exp28.py`, and `anonymous.py` (scopes J1–J4), respectively. Indexed receipts, including the four remaining-table public receipts, record hashes, execution environments, status/selection checks, and numerical comparisons. Internal release IDs T4, T5, and T10 refer to printed Tables 6, 7, and 9.

Public source includes the EXP04/06 runner dependency closures and the EXP28 controller, instrumentation, insertion, and restart wrappers. The remaining-table package provides replay code and saved evidence, not every original runner or historical artifact. Unreleased components remain private; Gurobi is proprietary. Xiong and colleagues' original-author RASstar source is linked rather than redistributed. Saved-output replay is distinct from independent optimizer execution and does not promise bit-identical stochastic reruns, cross-hardware runtime agreement, or exact-arithmetic solver proofs.

Funding. No funding was received for this research.

Competing interests. The author declares no competing interests.

Author contributions. Soroush Vahidi conceived the study; designed its methodology and experiments; developed the study-specific experiment and validation code; conducted the investigation, independent evidence checks, and analysis; and wrote and revised the manuscript. Externally authored algorithms and third-party software remain credited to their respective creators; advisor guidance and AI-assisted activities are acknowledged and disclosed separately. The author is responsible for the final content.

References

[1] 2F (2026) F-a-b-r-i-z-i-o/Lop-Time-For-a-Change: New benchmark suite for the linear ordering problem. Zenodo, version 0.0.1, https://doi.org/10.5281/zenodo.20554376, EXIOBASE-derived matrix archive. Registered creator: 2F; contact person: Fabrizio Fagiolo

[2] Bader DA, Ellis-Joyce J, Both GJ, et al (2025) Rocket-crane algorithm for the feedback arc set problem. Social Network Analysis and Mining 15(1):68. https://doi.org/10.1007/s13278-025-01491-2

[3] Baharev A, Schichl H, Neumaier A, et al (2021) An exact method for the minimum feedback arc set problem. ACM Journal of Experimental Algorithmics 26:Article 1.4. https://doi.org/10.1145/3446429

[4] Ben Ali A (2024) Comprehensive learning TLBO with recursive precedence-based solution construction and multilevel local search for the linear ordering problem. Expert Systems with Applications 238:122315. https://doi.org/10.1016/j.eswa.2023.122315

[5] Benito-Marimon M, Martí R, Martínez-Gavara A (2026) Multiple linear ordering problem. European Journal of Operational Research 331(2):351–364. https://doi.org/10.1016/j.ejor.2025.10.040

[6] Bollobás B, Borgs C, Chayes J, et al (2003) Directed scale-free graphs. In: Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp 132–139, URL https://ptolemy.berkeley.edu/projects/chess/pubs/724.html

[7] Cavallaro C, Cutello V (2025) An efficient heuristic algorithm to compute minimal and stable weighted feedback arc sets. In: Proceedings of the 37th International Conference on Software Engineering and Knowledge Engineering, pp 84–87, https://doi.org/10.18293/seke2025-049

[8] Cavallaro C, Cutello V, Pavone M (2024) Efficient heuristics to compute minimal and stable feedback arc sets. Journal of Combinatorial Optimization 48(4):30. https://doi.org/10.1007/s10878-024-01209-8

[9] Cavallaro C, Cutello V, Pavone M (2025) Minimal and stable feedback arc sets and graph centrality measures. Computers & Operations Research 184:107247. https://doi.org/10.1016/j.cor.2025.107247

[10] Dasdan A (n.d.) Directed graph benchmarks in DIMACS graph format. URL https://github.com/alidasdan/graph-benchmarks, study snapshot 77e5b8b4f1e1a1b8e80cfb9430f90492ffee3675; underlying origins vary

[11] Demetrescu C, Finocchi I (2003) Combinatorial algorithms for feedback problems in directed graphs. Information Processing Letters 86(3):129–136. https://doi.org/10.1016/S0020-0190(02)00491-X

[12] Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7:1–30. URL https://jmlr.org/papers/v7/demsar06a.html

[13] Dolan ED, Moré JJ (2002) Benchmarking optimization software with performance profiles. Mathematical Programming 91(2):201–213. https://doi.org/10.1007/s101070100263

[14] Eades P, Lin X, Smyth WF (1993) A fast and effective heuristic for the feedback arc set problem. Information Processing Letters 47(6):319–323. https://doi.org/10.1016/0020-0190(93)90079-O

[15] Erdős P, Rényi A (1959) On random graphs. I. Publicationes Mathematicae Debrecen 6:290–297. https://doi.org/10.5486/PMD.1959.6.3-4.12

[16] Even G, Naor J, Schieber B, et al (1998) Approximating minimum feedback sets and multicuts in directed graphs. Algorithmica 20(2):151–174. https://doi.org/10.1007/PL00009191

[17] Fagiolo F, Baioletti M, Santucci V (2027) Linear ordering problem: Time for a change. In: Parallel Problem Solving from Nature – PPSN XIX, Lecture Notes in Computer Science, vol 16985. Springer, p 399–416, https://doi.org/10.1007/978-3-032-36214-8_25, published online 25 August 2026; print volume dated 2027

[18] de Freitas CR, Luche JRD (2026) A parametrized synthetic benchmark and instance difficulty analysis for educational clinical internship scheduling with reciprocal supervision. International Journal of Data Science and Analytics 22:318. https://doi.org/10.1007/s41060-026-01316-1

[19] Grötschel M, Jünger M, Reinelt G (1984) A cutting plane algorithm for the linear ordering problem. Operations Research 32(6):1195–1220. https://doi.org/10.1287/opre.32.6.1195

[20] Gurobi Optimization, LLC (n.d.) Gurobi Optimizer reference manual. URL https://docs.gurobi.com/projects/optimizer/en/current/, study used Gurobi 12.0.3 with project-maintained SCC decomposition and lazy cuts

[21] Hecht M, Gonciarz K, Horvát S (2021) Tight localizations of feedback sets. ACM Journal of Experimental Algorithmics 26(1):Article 1.5. https://doi.org/10.1145/3447652

[22] Holland PW, Laskey KB, Leinhardt S (1983) Stochastic blockmodels: First steps. Social Networks 5(2):109–137. https://doi.org/10.1016/0378-8733(83)90021-7

[23] igraph Developers (n.d.) Finding a feedback arc set in a graph. URL https://r.igraph.org/reference/feedback_arc_set.html, documents Eades approximation and exact IP algorithm choices; study used python-igraph 1.0.0

[24] Internal Revenue Service (n.d.) SOI tax stats: Migration data 2021–2022. URL https://www.irs.gov/statistics/soi-tax-stats-migration-data-2021-2022, county-to-county outflow CSV

[25] Jean-Quartier C, Jeanquartier F, Stryeck S, et al (2025) Sharing practices of software artefacts and source code for reproducible research. International Journal of Data Science and Analytics 20(3):2539–2550. https://doi.org/10.1007/s41060-024-00617-7

[26] Kenyon-Mathieu C, Schudy W (2007) How to rank with few errors. In: Proceedings of the 39th Annual ACM Symposium on Theory of Computing. ACM, pp 95–103, https://doi.org/10.1145/1250790.1250806

[27] Leskovec J, Krevl A (2014) SNAP Datasets: Stanford large network dataset collection. URL https://snap.stanford.edu/data/, source files identified by content hashes in the study manifests

[28] Leskovec J, Huttenlocher D, Kleinberg J (2010) Signed networks in social media. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp 1361–1370, https://doi.org/10.1145/1753326.1753532

[29] Lugo L, Segura C, Miranda G (2022) A diversity-aware memetic algorithm for the linear ordering problem. Memetic Computing 14:395–409. https://doi.org/10.1007/s12293-022-00378-5

[30] Martí R, Reinelt G (2011) The Linear Ordering Problem: Exact and Heuristic Methods in Combinatorial Optimization, Applied Mathematical Sciences, vol 175. Springer, https://doi.org/10.1007/978-3-642-16729-4

[31] Martí R, Reinelt G, Duarte A (2012) A benchmark library and a comparison of heuristic methods for the linear ordering problem. Computational Optimization and Applications 51(3):1297–1317. https://doi.org/10.1007/s10589-010-9384-9

[32] NetworkX Developers (n.d.) NetworkX graph generators. URL https://networkx.org/documentation/stable/reference/generators.html, study's generated inputs used NetworkX 3.6.1 before project-specific normalization

[33] OpenFlights (n.d.) Airport and airline data. URL https://openflights.org/data, route and airport tables from jpatokal/openflights commit 56a3cce4d50309cb79948103a48c7f87d09b3056

[34] Panzarasa P, Opsahl T, Carley KM (2009) Patterns and dynamics of users' behavior and interaction: Network analysis of an online community.

Journal of the American Society for Information Science and Technology 60(5):911–932. https://doi.org/10.1002/asi.21015

[35] Pérez Lugo LJ, Segura C, Miranda G (2025) Future trends in the design of memetic algorithms: the case of the linear ordering problem. Neural Computing and Applications 37:12471–12485. https://doi.org/10.1007/s00521-025-11171-z

[36] Petelin G, Cenikj G (2025) The pitfalls of benchmarking in algorithm selection: What we are getting wrong. In: Proceedings of the Genetic and Evolutionary Computation Conference. ACM, pp 1181–1189, https://doi.org/10.1145/3712256.3726336

[37] Qian Y, Lin J, Li D, et al (2020) Block-insertion-based algorithms for the linear ordering problem. Computers & Operations Research 115:104861. https://doi.org/10.1016/j.cor.2019.104861

[38] Schiavinotto T, Stützle T (2004) The linear ordering problem: Instances, search space analysis and algorithms. Journal of Mathematical Modelling and Algorithms 3(4):367–402. https://doi.org/10.1023/B:JMMA.0000049426.06305.d8

[39] SciPy Developers (n.d.) scipy.optimize.milp: Mixed-integer linear programming. URL https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.milp.html, HiGHS-backed interface; study environment recorded SciPy 1.17.1

[40] Tarjan R (1972) Depth-first search and linear graph algorithms. SIAM Journal on Computing 1(2):146–160. https://doi.org/10.1137/0201010

[41] U.S. Census Bureau (n.d.) LEHD origin–destination employment statistics (LODES). URL https://lehd.ces.census.gov/data/lodes/, LODES8 Delaware 2022 main, all-jobs origin–destination file

[42] Vahidi S (2026) MWFAS benchmark extension: Metadata and objective witnesses. URL https://huggingface.co/datasets/SoroushVahidi/mwfas-benchmark-extension/tree/1e9db1daa5401b56398e9807d4e48d989d9f0eab, immutable public projection of audited experiment metadata; most original graph/order bytes are not included

[43] Vahidi S (2026) MWFAS heuristic metrics: Run-level outcomes for minimum weighted feedback arc set experiments. Hugging Face Datasets, URL https://huggingface.co/datasets/SoroushVahidi/mwfas-heuristic-metrics/tree/5b6cd5620ca0a1c9b3be5e8d996bf8e3c467e79f, fourteen-configuration snapshot; immutable revision in URL

[44] Vahidi S (2026) Training-free ranking from pairwise comparisons via acyclic graph construction. The Journal of Supercomputing 82:705. https://doi.org/10.1007/s11227-026-08852-4

[45] Vermetten D, Rook J, Preuß OL, et al (2025) MO-IOHinspector: Anytime benchmarking of multi-objective algorithms using IOHprofiler. In: Evolutionary Multi-Criterion Optimization, Lecture Notes in Computer Science, vol 15512. Springer, p 242–256, https://doi.org/10.1007/978-981-96-3506-1_17

[46] Wikimedia Foundation (n.d.) Wikipedia clickstream. URL https://dumps.wikimedia.org/other/clickstream/, monthly public dataset; exact pre-May-2026 source TSVs for the historical derived graphs are unavailable

[47] Xiong Z, Zhou Y, Xiao M, et al (2024) Finding small feedback arc sets on large graphs. Computers & Operations Research 169:106724. https://doi.org/10.1016/j.cor.2024.106724

[48] Yin H, Benson AR, Leskovec J, et al (2017) Local higher-order graph clustering. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 555–564, https://doi.org/10.1145/3097983.3098069