



Fixed Charges and Quantity Discounts in Unbalanced Transportation: Measured Limits of Exact Solving, an Amortized-Proxy Heuristic, and the Seed- Inheritance Anatomy of Ant Colony Optimization

Ali Mahmoud Assabri ^{1*}, Mohamed Al-Issawi Kajman ², Naji Abdalaziz Ali, ³

Abdulbasat Abdulghafar ⁴

alimahmoudassabri@gmail.com ✉

^{1,4} Department of Engineering Management, College of Technical Sciences, Bani Walid, Libya

² Department of Computing and Information Technology, College of Technical Sciences, Bani Walid, Libya

³ Department of Mechanical Engineering, College of Technical Sciences, Bani Walid, Libya

Article history	Received 12 January 2026	Accepted 16 March 2026	Publishing 10 April 2026
-----------------	-----------------------------	---------------------------	-----------------------------

ABSTRACT:

The unbalanced transportation problem rarely appears in practice with the symmetric penalties that textbooks assume: unshipped supply and unfilled demand carry heterogeneous costs, and carrier contracts add per-lane fixed charges and all-unit quantity discounts, which together render the model non-convex. What does the uniform-dummy convention actually cost, where does certified exact solving stop, what remains achievable in seconds, and what does a seeded ant colony contribute beyond its seed? We answer with one unified FCTP-QD model, one true-cost functional that scores every plan identically, six pipeline families, a certified MILP benchmark, and a seed-hierarchy ablation on instances from 3×4 to 200×200 . The convention pays 5.11–89.01% above the MILP incumbent at nominal fixed charges, and the heterogeneous linear remedy, exact in the purely linear case, breaks down by +14.35–60.37% once charges become realistic. Certified MILP gaps grow from 0% up to 20×20 (at most $\approx 1.1\%$ through 50×50) to 23.77% at 200×200 , and on the 200×200 discount instance the solver returns no feasible incumbent within its 120 s budget. The amortized fixed-charge proxy LP, by contrast, is numerically indistinguishable from the 120 s incumbent at 200×200 (+0.003%) in about 0.5 s, and on the 200×200 discount instance it is the only pipeline returning a competitive plan (76,859.52, against 79,839.69 for the next-best linear pipeline). The dissection is unambiguous: strong LP seeds pass through the colony unchanged (± 0), weaker seeds stall at their own basin, final quality rank equals seed rank, and unseeded colonies run +9.3–88.7% worse (Wilcoxon $p=9.77e-04$). Swarm quality here is inherited, not created. The result is a measured, honest benchmarking template — not an indictment of swarm methods, but a protocol for isolating what any mechanism adds.

Keywords: fixed-charge transportation; quantity discounts; unbalanced transportation problem; ant colony optimization; matheuristic seeding; amortized heuristic; computational benchmarking

1. Introduction

Transportation planning is the oldest optimization problem in logistics, yet two of its most routine cost features are still handled by convention rather than by measurement. This paper measures them. It prices the conventions, locates the point where exact solving stops, and dissects what a seeded swarm actually contributes.

1.1. Background and Motivation

The transportation problem (TP) asks how to ship a single commodity from sources I to destinations J at minimum linear cost, and it has anchored operations research since Hitchcock's formulation [1]. Real instances rarely balance. Total supply $\sum_{i \in I} a_i$ seldom equals total demand $\sum_{j \in J} b_j$, so textbooks append a dummy row or column that absorbs the imbalance at zero or uniform penalty cost [3], [4]. Practice is not uniform. Unshipped supply at source i incurs a holding cost h_i , unfilled demand at destination j triggers a shortage penalty p_j , and the two can differ across sites by an order of magnitude: warehousing surplus grain is not the same economic event as defaulting on a hospital's supply contract. A heterogeneous penalty-aware linear model captures this asymmetry exactly, and in the purely linear case it closes the convention's gap. The open question is what happens once the cost structure stops being linear. That question is not academic. Procurement plans built on the convention are executed under the true cost structure, so any modeling error surfaces later as realized cost — unbudgeted penalties, redundant truck hires, missed discount tiers.

Freight procurement supplies two nonlinear features, and they arrive together. Carriers quote per-lane fixed charges (truck hire, route opening, administrative setup), modeled here as f_{ij} , payable whenever arc i,j carries any positive flow. The same contracts reward consolidation: all-unit quantity discounts cut the per-unit rate once a shipment crosses tier thresholds, set in this study at $\tau_2 = 0.25$ and $\tau_3 = 0.5$ of $M_{ij} = \min(a_i, b_j)$, with discounted rates $r_2 = 0.85$ and $r_3 = 0.70$ [28]. Both mechanisms push toward fewer, heavier arcs, and both are standard line items in carrier contracts rather than exotic special cases. Both also destroy the total unimodularity that makes the plain TP solvable by inspection of its constraint matrix. The fixed-charge variant alone has been NP-hard since Balinski's formulation [24]; the all-unit discount adds a non-convex, discontinuous rate schedule on top, so the combined model resists every cheap trick that works for its linear ancestor.

Around this difficulty a pair of folklore claims has hardened. The first holds that LP-based and exact methods fail under such nonlinearity. The second holds that swarm intelligence excels precisely there [9], [14]. Neither claim, in this problem space, has been tested against certified dual bounds; both are repeated as justification rather than as result. The stakes are concrete: a planner choosing a tool at a given scale deserves to know what the cheap conventions cost, where the exact solver stops, and what the swarm adds beyond its initialization.

1.2. Research Gap and Problem Statement

The complexity status of the fixed-charge transportation problem (FCTP) is settled [24]; its practical status is not. Four specific gaps motivate this study. First, the cost of the uniform-dummy convention has never been measured under fixed charges and quantity discounts with heterogeneous penalties. The convention survives in textbooks and software defaults, yet no published account prices its premium against a certified benchmark on instances where the penalties differ across sites. Second, the practical ceiling of modern MILP branch-and-bound on FCTP and FCTP-QD is unmeasured. NP-hardness says nothing about where a contemporary dual-bounding solver stalls; the boundary between "exact in seconds" and "no certified answer" is empirically unknown for this model class. Third, no seconds-fast, interpretable heuristic is documented at the 200×200 scale (40,000 arcs, and 160,000 binaries once discount tiers are expanded) with its quality stated against certified bounds rather than against other heuristics. Fourth, and consequential in a way the literature overlooks, swarm studies in this space report final objective values without seed ablations. When an ant colony is warm-started from a strong construction heuristic, the published result confounds two contributions: the seed's and the swarm's. Published FCTP metaheuristics [26], [27] follow exactly this pattern, as does the broader swarm-for-transportation literature surveyed in [16]. A colony that returns an excellent plan may have discovered it, inherited it, or merely polished it; without ablating the seed, the three stories are indistinguishable, and the marginal value of the swarm mechanism itself has never been isolated.

We address all four gaps with one experimental design, built so that every claim reduces to a number. The object of study is a unified FCTP-QD model: flow variables $x_{ij} \geq 0$, arc-opening variables $z_{ij} \in \{0,1\}$, unit costs c_{ij} , fixed charges f_{ij} , supplies a_i , demands b_j , holding costs h_i , shortage penalties p_j , and a tier function $\gamma(x)$ that applies the rate 1, r_2 , or r_3 according to the share of M_{ij} shipped. Every plan, from whatever pipeline, is scored by one true-cost functional (tiered transport cost plus fixed charges plus heterogeneous penalties), so that comparisons are apples-to-apples even when a pipeline optimizes a surrogate. Six pipelines are run on nine instances I1–I9 spanning 3×4 to 200×200 : Dummy (the uniform convention), Remedy (the heterogeneous linear model that ignores charges and tiers), MILP (certified branch-and-bound under a 120 s budget), Proxy (the amortized fixed-charge LP we propose), its LNS and basin-hopping polishes (Proxy+LNS, Proxy+BH), and seeded and unseeded ant colony optimization (ACO, ACO-NS; $\alpha=1$, $\beta=2$, $\rho=0.5$). The instance ladder is not decorative. Small sizes admit exact optima and expose structural error; mid sizes test whether heuristics track a converging solver; the 200×200 frontier tests who still answers when certification is out of reach. One model, one scoring functional, six pipelines, nine sizes: the design is deliberately plain so that the numbers carry the argument.

1.3. Contribution Statement

This paper makes five contributions, each tied to a measured number.

1. **The first measurement of the uniform-dummy convention's premium under heterogeneous penalties and nonlinear charges.** Evaluated under the true-cost functional, the convention pays 5.11–89.01% above the MILP incumbent across I1–I9 at nominal fixed charges, compounding with charge magnitude and scale. The heterogeneous linear remedy, exact in the linear case, breaks down once fixed charges become realistic: its certified-incumbent gaps run +14.35–60.37% under the realistic charge regime. The convention and its remedy both fail, and we show where.
2. **A certified MILP scalability ceiling for FCTP-QD.** Rather than asserting solver failure, we measure it: certified gaps of 0% up to 20×20 , and at most $\approx 1.1\%$ through 50×50 , degrade to 8.67% under nominal charges and 23.77% under realistic charges at 200×200 , and on the 200×200 quantity-discount instance the solver returns no feasible incumbent within the 120 s solver budget (≈ 140 s wall incl. ≈ 19 s build; 160,000 binaries). Frontier claims throughout are incumbent comparisons, not optimality claims; at 200×200 nothing is certified near-optimal.
3. **The amortized fixed-charge proxy LP:** a seconds-fast, interpretable heuristic. Spreading each fixed charge over M_{ij} and discounting variable costs by $r_{\{2\}}$ converts the nonlinear problem into one standard unbalanced TP LP. The resulting plans sit within 0–5.3% of the MILP incumbent at nominal charges up to 50×50 , and at 200×200 they are numerically indistinguishable from the 120 s MILP incumbent, at $\sim 0.5\%$ (+0.003% nominal, +0.01% realistic). On the 200×200 discount instance, where the MILP returns no feasible incumbent within the 120 s solver budget, the proxy is the only pipeline returning a competitive plan: 76,859.52

in seconds, against 79,839.69 for the next-best linear pipeline. Optional arc-deletion LNS and basin-hopping polish shave a further -0.1% to -3.9% .

4. **The seed-inheritance anatomy of ant colony optimization.** Dual-seeded with LP plans, the colony returns its seed unchanged on every tested instance (± 0 , $\text{std} = 0$): zero marginal contribution. Seeded with weaker textbook constructions, the swarm cuts $45.3\text{--}79.3\%$ off the raw seed's cost yet stalls at the basin of those heuristics, never reaching the LP basin; the final quality rank equals the seed rank. Unseeded colonies run $+9.3\text{--}88.7\%$ worse than seeded ones (Wilcoxon $p=9.77e-04$). What the swarm returns was present at initialization — a measurement, not a conjecture, and a caution for how seeded swarm results are reported.
5. **An honest-benchmarking protocol as a reusable template.** One true-cost functional for all pipelines, certified dual bounds wherever the solver provides them, and frontier statements restricted to incumbent comparisons. The protocol converts unverifiable claims, of solver failure or of swarm superiority, into numbers that can be checked, and it transfers unchanged to other problem classes.

1.4. Paper Organization

Section 2 positions the study against prior work on unbalanced transportation, fixed-charge models, quantity discounts, and ant colony methods, and tabulates which of these carry certified benchmarks, convention measurements, and seed ablations — few carry any of the three. Section 3 states the unified FCTP-QD formulation, its computational status, and the MILP benchmark with its binary accounting. Section 4 details the six pipelines and the evaluation protocol: the dummy and remedy conventions, the HiGHS MILP configuration, the amortized-proxy LP with its LNS and basin-hopping polish, the seeded ACO with its seed-hierarchy ablation, and the statistical machinery. Section 5 reports the results — the measured MILP ceiling, the convention premium and remedy breakdown, the proxy's frontier behavior, and the seed-inheritance anatomy — and Section 6 interprets the three scale regimes, explains why the proxy survives where the remedy fails, and lists the limitations. Section 7 concludes with the contributions restated in numbers and four directions for future work. Appendix A gives the instance data and generation protocol for full reproducibility.

2. Literature Review

2.1. From Hitchcock to Heterogeneous Penalties

The transportation problem (TP) enters the literature with Hitchcock [1], who asked how to distribute a homogeneous product from several sources to numerous localities at least cost, and Koopmans [2], who gave the model its activity-analysis reading. Both treated the balanced case, in which aggregate supply equals aggregate demand, as canonical. Their formulations also carry the property that drives the classical theory: the constraint matrix of a TP is totally unimodular, so integer data yield integer optima and the problem reduces to an easy linear program (LP).

Real networks are rarely balanced. Sources hold surplus; destinations absorb shortages. The textbook fix, codified in Hillier and Lieberman [3] and Taha [4], appends a dummy destination to absorb excess supply or a dummy source to cover excess demand, and prices every dummy arc at one uniform rate, frequently zero. The convention is tidy. It also assumes, silently, that holding a unit at source i costs the same for all i , and that a unit short at destination j hurts the same for all j . A parallel strand relaxed the balance requirement structurally rather than cosmetically: Brigden [5] analyzed transportation problems whose constraints are of mixed type, showing that imbalance can be absorbed inside an augmented tableau without the dummy fiction. Together these works establish that the balanced-with-uniform-dummy model is a convention, not a theorem. What none of them measures is the price of that convention when the true penalties (a per-source holding cost h_i on excess supply and a per-destination shortage penalty p_j on excess demand) are heterogeneous, and when the cost structure is no longer linear. In the purely linear case, a companion study by the present authors measured that premium and showed a penalty-priced heterogeneous dummy to be an exact remedy. What happens once nonlinearity enters is, to our knowledge, unmeasured anywhere.

The first nonlinearity is the fixed charge. Attaching a cost f_{ij} to the mere act of opening arc i, j , independent of the flow x_{ij} it carries, produces the fixed-charge transportation problem (FCTP). Balinski formulated the FCTP as a mixed-integer program and proved it NP-hard [24]. The reason is structural: the arc-open decisions $z_{ij} \in \{0,1\}$ destroy total unimodularity, and the objective jumps discontinuously at zero flow. Everything the classical apparatus guarantees (exactness in milliseconds, integer optima from a continuous solve) vanishes with a single parameter.

2.2. Exact and Heuristic Methods for FCTP

Balinski's formulation [24] opened an exact line: branch-and-bound over the arc-open binaries, with linear relaxations supplying dual bounds. The line is rigorous—every certified FCTP optimum in the literature traces back to it or to a descendant—but its practical ceiling has rarely been published with certified gaps attached. Reported experiments typically stop where the solver still succeeds, so the boundary between “solvable” and “out of reach” remains anecdotal rather than mapped.

Heuristics filled the gap early. Adlakha and Kowalski [25] proposed a simple heuristic for small fixed-charge problems, a method that earns its keep in teaching and in rapid what-if analysis but carries no certificate of quality. Metaheuristics followed: a spanning-tree genetic algorithm [26] and a Prüfer-number-encoded immune/genetic hybrid [27] attacked the FCTP with population methods, and hybrid meta-heuristics have since been built for a multi-route capacitated variant [20]. These studies report competitive objective values. Still, three features limit their evidentiary weight for the questions asked here. Comparisons usually run against other metaheuristics, not against a certified dual bound. Instance sizes top out well below modern distribution scale. And none ablates its initialization, so the reader cannot tell how much of the reported quality belongs to the search and how much arrived with the starting population.

Modern mixed-integer technology complicates this picture. Contemporary branch-and-cut codes solve structured mixed-integer linear programs (MILPs) that would have been unthinkable in 1961, and the open-source HiGHS solver puts that

machinery within reach of the FCTP without license barriers. Where its practical ceiling sits on fixed-charge problems of realistic size (100×100, 200×200, with tens or hundreds of thousands of binaries) has not, as far as we know, been measured against certified dual bounds. We therefore treat MILP as the benchmark to approach or to excuse, and we measure its ceiling rather than assume it.

2.3. Quantity Discounts in Transportation

The second nonlinearity is the quantity discount. Balachandran and Perry [28] introduced the TP with all-unit discounts, in which the unit rate on an arc depends on the entire quantity shipped there: cross a threshold and the cheaper rate applies retroactively to every unit. The distinction from incremental discounts matters computationally. Incremental schemes keep total cost piecewise linear and concave; all-unit schemes do neither. At a tier boundary the cost drops discontinuously, the surface is non-convex, and marginal reasoning fails exactly where the decisions are made.

We model the all-unit case with a tier function $\gamma(x)$ defined relative to the arc capacity $M_{ij} = \min(a_i, b_j)$, where a_i is the supply at source i and b_j the demand at destination j : $\gamma(x) = 1$ if $x < \tau_2 M_{ij}$; $\gamma(x) = r_2$ if $\tau_2 M_{ij} \leq x < \tau_3 M_{ij}$; and $\gamma(x) = r_3$ if $x \geq \tau_3 M_{ij}$, with $\tau_2 = 0.25$, $\tau_3 = 0.5$, $r_2 = 0.85$, and $r_3 = 0.70$ held fixed throughout. The discount rewards consolidation (fewer open arcs, heavier flows), which is precisely the opposite pressure to the fixed charge's penalty on arc opening. An FCTP with all-unit quantity discounts (FCTP-QD) therefore contains an internal tension: open few arcs to save charges, but open enough of them to place flows where variable costs are low.

The literature at this exact intersection is thin. Balinski's complexity result [24] covers the fixed charge; Balachandran and Perry [28] cover the discount; heuristic and metaheuristic work treats one nonlinearity at a time. We are aware of no study that measures exact-solver behavior, heuristic quality, and swarm behavior jointly on the combined model. That thinness is not a gap inflated for rhetorical effect. It is half the motivation for this paper.

2.4. Swarm Intelligence for Transportation Problems

Ant colony optimization (ACO) grew from Dorigo's doctoral work [8] into a mature family [9] that constructs solutions probabilistically from pheromone memory and heuristic desirability, alongside the genetic-algorithm [11] and particle-swarm [12] lineages; standard ACO practice fixes the balance between the two with parameters α , β , and an evaporation rate ρ . Transportation engineering adopted swarm ideas early: Teodorović [14] reviewed the field and argued that swarm methods suit large, dynamic, or poorly structured problems; hybrid swarm metaheuristics have also been compared in supply-chain network design [23]. Applications directly on the TP are scarcer; Pandian and Natarajan [6] treat the mixed-constraint TP. Ekanayake et al. [22] modified the ACO transition rule for the classical TP and reported competitive results on benchmark sets—competitive against other heuristics, that is, not against a certified optimum on an identical objective. Jovanović et al. [13] applied ACO to graph partitioning under supply–demand constraints, a structural relative of the TP. A recurring empirical pattern, visible in vehicle-routing work such as [19] and applied ACO studies such as [18], is that ACO earns its best results when paired with local search; the bare construction mechanism rarely suffices on its own, consistent with broader metaheuristic taxonomies for routing [21].

Initialization is the quiet variable in all of this. The classical construction heuristics—the northwest-corner rule (NW), the least-cost method (LCM), and Vogel's approximation (VAM)—appear in every textbook as starting bases for exact procedures [3], [4], and improved initial solutions remain an active research topic [7]. Metaheuristic FCTP studies likewise seed their populations with structured encodings rather than random plans [26], [27]. What the literature does not isolate is the seed's causal contribution. If a swarm seeded from a strong LP-derived plan reports a good final cost, how much of that cost was already present at iteration zero? The no-free-lunch theorems [15] warn that no search dominates universally; they say nothing about attribution. Prior ACO-TP results may well be correct as far as they go. We do not dispute them. The attribution question is simply separate: whether the colony improves on its seed or merely inherits it cannot be answered without seed-source ablations, and the published studies do not report any. That unmeasured interaction, not any suspected error in prior work, is what our seed-hierarchy protocol is designed to expose.

2.5. Positioning

Four unmeasured quantities define this paper. (i) The uniform-dummy convention's cost is unmeasured once penalties are heterogeneous and fixed charges are realistic; the linear-case measurement exists, the nonlinear one does not. (ii) The practical ceiling of MILP on FCTP-QD at 200×200 has no published, certified measurement; in our own tier-expanded benchmark the solver returns no feasible incumbent within the 120 s solver budget (160,000 binaries) on the 200×200 QD instance. (iii) No seconds-fast, interpretable heuristic has been evaluated at that frontier; our amortized-proxy LP, which spreads each fixed charge f_{ij} over the capacity M_{ij} and solves a single transportation LP, is designed to fill that slot. (iv) The seed-inheritance question (how ACO interacts with its initialization) has no anatomy in the literature. Table 1 places the closest literature threads against these four gaps.

Table 1. Positioning against the closest literature threads.

Closest thread	What it shows	What remains unmeasured
Unbalanced-TP conventions and remedies [3]–[5]	Balancing is a modeling choice; the heterogeneous-penalty linear case is exactly solvable	The convention's cost and the remedy's breakdown once fixed charges and discounts enter
FCTP exact and metaheuristic methods [24]–[27], [20]	NP-hardness; small-instance heuristics; genetic and immune designs reporting competitive values	Certified MILP ceiling at 200×200 ; evaluation against dual bounds; seed ablations

ACO for transportation [8], [14], [22]	Probabilistic construction suits network problems; benchmark-competitive results	Marginal contribution over the seed; performance against certified bounds at scale
---	---	--

The middle column is generous: each thread genuinely shows something, and none of the four gaps implies that prior results are wrong. The right-hand column is the point. Reading down it, a pattern emerges: everything a practitioner would need in order to choose a method rationally is unmeasured: the cost of the convenient formulation, the reach of the exact one, the existence of a fast interpretable option in between, and the true source of swarm-reported quality. Reading across, the threads are complementary rather than competing; each stops one step short of the measurement taken here. Two of the three rows converge on the same missing discipline (certified baselines and seed attribution), which is why our protocol builds both in from the start instead of appending them as robustness checks. The model of Section 3 unifies the nonlinearities these threads treat separately, and the protocol of Section 4 turns the table's empty cells into measured ones.

3. Mathematical Model

This section states the unified fixed-charge transportation problem with quantity discounts (FCTP-QD) that the remainder of the paper solves, benchmarks, and stress-tests. The formulation is conservative by design: every element is standard on its own, and the difficulty comes from their interaction. We first fix sets and parameters, then give the model, then show that it nests the classical problems it extends, and finally discuss what its cost geometry implies for exact solution.

3.1. Problem Setting and Notation

Let I be the set of sources, $|I| = m$, and J the set of destinations, $|J| = n$. Source $i \in I$ holds a supply $a_i > 0$; destination $j \in J$ requires a demand $b_j > 0$. We do not assume balance. Throughout, the instance is in the excess-supply (ES) regime when $\sum_{i \in I} a_i > \sum_{j \in J} b_j$ and in the excess-demand (ED) regime when the inequality is reversed; equality is the degenerate balanced case. Imbalance is not free. Supply that cannot be shipped is held at source i at a unit holding cost $h_i \geq 0$, and demand that cannot be served incurs a unit shortage penalty $p_j \geq 0$. Both rates are heterogeneous: they vary across sources and destinations, so where the imbalance falls matters as much as how large it is. This heterogeneity is what the uniform-dummy convention of §3.5 discards.

Moving one unit from i to j costs $c_{ij} \geq 0$. Using the arc at all costs $f_{ij} \geq 0$, a fixed charge paid once if arc (i, j) carries any positive flow and zero otherwise. Fixed charges model consolidation economies (a truck hired, a lane contracted, a pipeline segment activated), and they push optimal plans toward few, heavily loaded arcs.

The third cost element is an all-unit quantity discount on the variable rate, in the tradition of Balachandran and Perry [28]. Define $M_{ij} = \min(a_i, b_j)$, the largest flow arc (i, j) can ever carry; discount tiers are expressed relative to this bound so that the schedule is comparable across arcs of different capacities. With tier thresholds $\tau_2 = 0.25$ and $\tau_3 = 0.5$ of M_{ij} and tier rates $r_2 = 0.85$ and $r_3 = 0.70$, the effective unit-rate multiplier on arc (i, j) is

$$\gamma(x) = \begin{cases} 1, & x < \tau_2 M_{ij}, \\ r_2, & \tau_2 M_{ij} \leq x < \tau_3 M_{ij}, \\ r_3, & x \geq \tau_3 M_{ij}. \end{cases}$$

The discount is all-unit: once flow crosses a threshold, the cheaper rate applies to every unit on the arc, not only to units beyond the threshold. A consequence worth pausing on is that more freight can cost less money. A flow just below $\tau_2 M_{ij}$ pays c_{ij} per unit on all of it; a flow exactly at $\tau_2 M_{ij}$ pays $0.85 c_{ij}$ on all of it. The discount therefore rewards the same consolidation behavior as the fixed charge, and the two mechanisms compound rather than compete.

3.2. The Unified FCTP-QD Formulation

Two decision variables describe the transport plan: $x_{ij} \geq 0$, the flow on arc (i, j) , and $z_{ij} \in \{0, 1\}$, which takes the value 1 when arc (i, j) is open. Two slack variables describe the imbalance: $e_i \geq 0$, the surplus held at source i , and $u_j \geq 0$, the demand left unmet at destination j . The model minimizes total true cost,

$$\min \Phi = \sum_{i \in I} \sum_{j \in J} [\gamma(x_{ij}) c_{ij} x_{ij} + f_{ij} z_{ij}] + \sum_{i \in I} h_i e_i + \sum_{j \in J} p_j u_j,$$

subject to flow conservation, arc linking, and domains. The objective separates into three economically distinct parts. The first term prices movement, at a unit rate that depends on the arc's own flow through γ . The second prices the decision to use an arc at all. The third and fourth price the imbalance where it physically falls. Every pipeline studied in §4 can be characterized by which of these parts it treats exactly and which it approximates: the linear remedy keeps only the movement and penalty terms, while the proposed proxy rewrites the fixed charge as an amortized per-unit rate.

Conservation depends on the imbalance regime. Under excess supply, every demand can be served and the surplus stays at the sources:

$$\sum_{j \in J} x_{ij} + e_i = a_i, \quad i \in I; \quad \sum_{i \in I} x_{ij} = b_j, \quad j \in J.$$

Under excess demand, every supply is shipped and part of the demand goes unfilled:

$$\sum_{j \in J} x_{ij} = a_i, \quad i \in I; \quad \sum_{i \in I} x_{ij} + u_j = b_j, \quad j \in J.$$

One may equivalently write a single unified system, $\sum_j x_{ij} + e_i = a_i$ and $\sum_i x_{ij} + u_j = b_j$ with both slacks present; whenever $h_i > 0$ and $p_j > 0$, the slack that is structurally unnecessary vanishes at any optimum, so (3) and (4) are recovered regime by regime. We keep the two cases explicit because the pipelines in §4 treat them differently.

The fixed charge is activated by a standard linking inequality together with the variable domains:

$$x_{ij} \leq M_{ij} z_{ij}, \quad (i, j) \in I \times J; \quad x_{ij} \geq 0, \quad z_{ij} \in \{0,1\}, \quad e_i \geq 0, \quad u_j \geq 0.$$

The linking bound is tight by construction. No feasible flow exceeds $M_{ij} = \min(a_i, b_j)$, so $z_{ij} = 1$ never constrains the plan, while $z_{ij} = 0$ forces $x_{ij} = 0$ and exempts the arc from its fixed charge. Nothing in (2) – (5) is exotic. Yet the objective is nonlinear in x through γ , and combinatorial through z ; the model sits exactly at the point where linear transportation theory stops applying.

3.3. Nesting Property

Remark. The FCTP-QD nests its three ancestors, and the nesting chain localizes where tractability is lost.

Setting $r_2 = r_3 = 1$ makes $\gamma(x) \equiv 1$ and removes the discount; what remains is the fixed-charge transportation problem (FCTP) of Balinski [24], augmented with imbalance penalties. Setting additionally $f_{ij} = 0$ for every arc removes the combinatorial structure entirely: the model collapses to a linear, penalty-aware unbalanced transportation problem whose constraint matrix is totally unimodular, so its LP relaxation is exact and yields integral flows for integral data. If, further, $\sum_i a_i = \sum_j b_j$, the slacks e_i and u_j play no role and the classical balanced transportation problem of Hitchcock [1] is recovered. In symbols,

FCTP-QD $\supset$ FCTP $\supset$ linear unbalanced TP $\supset$ balanced TP.

Read left to right, each containment adds one difficulty. The balanced TP is solvable by inspection-grade LP machinery. Fixed charges destroy total unimodularity and introduce genuine combinatorics. Discounts then break the linearity of the arc cost itself. The pipelines evaluated in §5 can be read against this chain: the linear remedy operates at the third level and is applied to problems at the first.

3.4. Computational Status

The FCTP-QD is NP-hard. The argument needs no new machinery: the balanced, penalty-free instances with $r_2 = r_3 = 1$ are precisely the FCTP, which has been NP-hard since Balinski's formulation [24], so NP-hardness holds by restriction even before discounts enter. Historically, this status pushed the literature toward heuristics already for small instances [25]; §5 measures what a modern branch-and-bound solver actually achieves.

The discount term adds a distinct, geometric difficulty that NP-hardness alone does not capture. The arc cost $g_{ij}(x) = \gamma(x) c_{ij} x$ is piecewise linear with decreasing slopes $c_{ij} > r_2 c_{ij} > r_3 c_{ij}$, but it is also discontinuous: at $x = \tau_2 M_{ij}$ the cost jumps down by $(1 - r_2) c_{ij} \tau_2 M_{ij} = 0.0375 c_{ij} M_{ij}$, and at $x = \tau_3 M_{ij}$ by a further $(r_2 - r_3) c_{ij} \tau_3 M_{ij} = 0.075 c_{ij} M_{ij}$. The function is therefore neither convex nor concave, and its marginal cost vanishes precisely at the tier boundaries. No purely continuous LP reformulation exists; integrality is unavoidable.

For the certified benchmark we expand each arc into its three tiers. Introduce tier-selection binaries $y_{ij}^k \in \{0,1\}$ and tier flows $x_{ij}^k \geq 0$ for $k \in 1,2,3$, with breakpoints $T_{ij}^0 = 0, T_{ij}^1 = \tau_2 M_{ij}, T_{ij}^2 = \tau_3 M_{ij}, T_{ij}^3 = M_{ij}$ and rates $r_1 = 1, r_2, r_3$. The expansion enforces exactly one active tier per open arc and confines the flow to that tier's interval:

$$x_{ij} = \sum_{k=1}^3 x_{ij}^k, \quad \sum_{k=1}^3 y_{ij}^k = z_{ij}, \quad T_{ij}^{k-1} y_{ij}^k \leq x_{ij}^k \leq T_{ij}^k y_{ij}^k, \quad k \in \{1,2,3\},$$

and the arc term in (2) is replaced by its exact linearization

$$\sum_{k=1}^3 r_k c_{ij} x_{ij}^k + f_{ij} z_{ij}.$$

One technicality deserves mention. The half-open intervals of (1) become closed in (6), so the breakpoints belong to two tiers at once; the ambiguity is harmless because a minimizer always selects the cheaper tier at a shared breakpoint, making the expansion exact at optima.

The binary accounting is unforgiving. A plain FCTP on an $m \times n$ instance carries $m \cdot n$ binaries: 2,500 at 50×50 and 40,000 at 200×200 . The tier expansion adds three binaries per arc on top of the arc-open variable, giving $4 \cdot m \cdot n$ in our implementation: 100 at 5×5 , 10,000 at 50×50 , and 160,000 at 200×200 , with the continuous flow variables tripling alongside. Because every arc of the $m \cdot n$ grid must carry its own tier logic, the growth is quadratic in the side length of a square instance — a fourfold scale-up from 50×50 to 200×200 multiplies the binary count sixteenfold. Section 5.1 records the consequence: certified gaps widen with scale and, at 200×200 with discounts, the solver returns no feasible incumbent within the 120 s budget. The model of (6) – (7) is exact; it is simply not always usable.

3.5. The Uniform-Dummy Convention and Its Cost

Classical textbook treatments [3], [4] handle imbalance by balancing the table. The heterogeneous penalty structure is collapsed into uniform rates, typically the means $\hat{h} = \frac{1}{m} \sum_{i \in I} h_i$ and $\hat{p} = \frac{1}{n} \sum_{j \in J} p_j$; a dummy destination with demand $\sum_i a_i - \sum_j b_j$ and unit cost $\hat{h}$ is appended in the ES regime, or a dummy source with supply $\sum_j b_j - \sum_i a_i$ and unit cost $\hat{p}$ in the ED regime. The resulting balanced LP is solved, and the plan is then evaluated under the true heterogeneous costs. The convention is convenient, and in the purely linear setting its damage is bounded. Under fixed charges and discounts it loses the ability to decide where surplus should be held and which demand should go unmet — precisely the decisions the heterogeneity encodes. Section 5.2 measures the premium this simplification pays.

4. Methods

4.1. Overview: One True-Cost Functional, Six Pipeline Families

Every method in this study answers the same question, and every answer is graded on the same scale. Each pipeline receives an FCTP-QD instance — a set I of m sources, a set J of n destinations, unit transport costs c_{ij} , fixed charges f_{ij} , supplies a_i , demands b_j , and the heterogeneous imbalance penalties h_i (holding, excess supply) or p_j (shortage, excess

demand) — and returns a flow plan $x_{ij} \geq 0$ with implied arc-open indicators $z_{ij} = 1[x_{ij} > 0]$. The plan is then scored exactly once, under the true-cost functional of §3: the tiered transport term $\gamma(x_{ij}) c_{ij} x_{ij}$, with γ the all-unit discount multiplier defined there, plus the fixed charges $f_{ij} z_{ij}$, plus the penalty terms for residual imbalance. No pipeline is graded on its internal surrogate. A method may optimize whatever approximation it likes; it is judged by the tiered, fixed-charge, penalty-laden cost it actually incurs. This single-evaluator discipline is what makes the numbers of §5 comparable at all, and it is precisely what the “solver failure” folklore in the literature lacks.

Six pipeline families, comprising eight configurations, pass through this evaluator. They span three methodological generations: textbook linear practice, modern exact branch-and-bound, and metaheuristic search. Table 2 summarizes the roster before the subsections detail each entry.

Table 2. Pipelines evaluated under the common true-cost functional. “LP” denotes one solve of the unbalanced transportation linear program; MILP is HiGHS branch-and-bound via `scipy.optimize.milp`.

Pipeline	Type	Guarantee	Solver / engine	Budget
Dummy	linear convention (uniform dummy)	none	one LP	seconds
Remedy	penalty-aware linear heuristic	exact only when $F = 0$ and tiers flat	one LP	seconds
MILP	exact branch-and-bound (tier expansion of §3.4)	certified dual bound; gap reported	HiGHS via <code>scipy.optimize.milp</code>	120 s solver budget
Proxy	amortized fixed-charge LP heuristic (proposed)	none; interpretable surrogate	one LP	$0.5sat200 \times 200$
Proxy+LNS	proxy + arc-deletion large-neighborhood search	monotone true-cost improvement	$k = 15$ restricted LP re-solves, $3passessecondsProxy + BHproxy + basinhoppingmonotonetrue - costimprovement20kicks; LNSrepair(k = 25, 5 \text{ passes})$	seconds to minutes
ACO	dual-seeded ant colony analyzer	never worse than its best seed (by construction)	$30ants,$ 100 iterations	$25.7sperrunat200 \times 200QD$
ACO-NS	unseeded ant colony baseline	none	identical to ACO, no plan seed	identical

Two properties of the roster deserve emphasis before the details. First, the set is asymmetric in what it can certify. Only MILP produces a dual bound, so only MILP rows can carry the word “exact”; every other entry is an incumbent generator, and at the 200×200 frontier all comparisons in §5 are incumbent comparisons, not optimality claims. Second, the Proxy family is the methodological center of gravity. Dummy and Remedy exist to measure established linear practice; MILP exists to draw the certified ceiling; ACO and ACO-NS exist as analyzers whose seed-inheritance anatomy is itself a result. The proposed contribution is the amortized proxy and its two optional polishers. The asymmetry is intentional: a paper about measured limits needs baselines that measure, not baselines that merely compete.

4.2. The Dummy Pipeline (Convention Baseline)

The dummy pipeline reproduces the textbook balancing convention [3], [4]. When total supply exceeds total demand, a fictitious destination is appended; when demand exceeds supply, a fictitious source. The convention must price the dummy arcs, and the standard choice is a uniform unit cost, here the mean of the true heterogeneous penalties, $\hat{h} = \frac{1}{m} \sum_{i \in I} h_i$ or $\hat{p} = \frac{1}{n} \sum_{j \in J} p_j$, depending on the imbalance direction. The instance is then balanced and reduces to Hitchcock’s classical transportation problem [1], which a single linear-programming solve settles. Dummy arcs are discarded from the solution, the residual imbalance is mapped back to the real sources or destinations, and the recovered plan is re-evaluated under the true-cost functional with the actual h_i and p_j .

Two blind spots are structural. The uniform penalty flattens genuine heterogeneity: a source with holding cost 2 and one with 15 become indistinguishable. And the nonlinear structure (fixed charges f_{ij} and discount tiers) never enters the optimization at all, since the dummy LP sees only c_{ij} and the averaged penalty. The pipeline therefore measures what the most widely taught convention costs when the cost surface is no longer linear.

4.3. The Remedy Pipeline

The remedy pipeline repairs the first blind spot and only the first. Rather than a uniform dummy, each unit of excess supply at source i is charged its true holding cost h_i , and each unit of unmet demand at destination j its true shortage penalty p_j ; imbalance becomes a set of explicit slack decisions priced individually. The resulting model is still a single linear program over the standard unbalanced transportation constraints. It is solved once, and the plan is re-scored under the true-cost functional exactly as before.

This pipeline is the faithful version of linear practice, and it has a useful exactness property: when $f_{ij} = 0$ on every arc and the discount rates are flat ($r_2 = r_3 = 1$), the remedy LP is the true problem, by the nesting property of §3.3. Its only remaining blind spot is the nonlinearity itself: the fixed charges and tier jumps are absent from its objective. Where the remedy degrades in §5, the cause can therefore be attributed to that omission alone, not to penalty misspecification. This makes the remedy the cleanest possible measuring stick for the cost of ignoring f_{ij} .

4.4. MILP Benchmark

The exact benchmark is the tier-expanded mixed-integer formulation of §3.4: each arc's flow is decomposed into tier segments with associated binaries, the arc-open indicators z_{ij} couple flow to fixed charges, and the resulting program is linear in its expanded variables. Size accounting matters here. A 50×50 QD instance carries 10,000 binaries; the 200×200 QD instance carries 160,000. We solve with HiGHS branch-and-bound through `scipy.optimize.milp`, under a 120 s solver budget per instance; model-build time is wall-clocked separately and reported, never folded into the solver budget.

Reporting follows a strict gap discipline. A result is labeled “exact” only when the solver certifies optimality within the budget. Otherwise we report the incumbent, the certified dual bound, and the gap, computed as $(\text{incumbent} - \text{dual}) / \text{incumbent}$. One case requires exact phrasing, because it is the empirical replacement for the folklore claim that “exact methods fail”: on the 200×200 QD instance, HiGHS returns no feasible incumbent within the 120 s solver budget ($\approx 140s_{\text{wallincl.}} \approx 19s_{\text{build}}; 160,000 \text{ binaries}$). We state this as a measurement of one solver under one budget. Commercial solvers were not tested; nothing here generalizes to them (§6.2).

4.5. The Proposed Amortized-Proxy Heuristic

The proposed method rests on one observation about fixed charges. The charge f_{ij} is paid once if arc (i, j) carries any positive flow, regardless of how much. Ex ante, the fairest way to price that liability per unit is to spread it over the largest shipment the arc could plausibly sustain, which is $M_{ij} = \min(a_i, b_j)$: no arc can ever carry more than the smaller of its endpoints' budgets. Each arc then acquires an amortized unit cost $c_{ij} + f_{ij}/M_{ij}$. The quantity has a plain reading: it is the per-unit break-even rate of the arc, the effective unit cost at which shipping a full M_{ij} exactly absorbs the fixed charge. On quantity-discount instances, where an open arc will typically operate in a discounted tier, we additionally anticipate the middle-tier multiplier, replacing c_{ij} by $r_2 c_{ij}$ with $r_2 = 0.85$.

The heuristic solves a single penalty-aware linear program in these amortized costs:

$$\min_{x \geq 0, e \geq 0, u \geq 0} \sum_{i \in I} \sum_{j \in J} \left(\delta_{ij} c_{ij} + \frac{f_{ij}}{M_{ij}} \right) x_{ij} + \sum_{i \in I} h_i e_i + \sum_{j \in J} p_j u_j,$$

subject to the unbalanced transportation constraints of §3.2, where e_i is the excess-supply slack at source i , u_j the unmet-demand slack at destination j , and $\delta_{ij} = r_2$ on QD instances and $\delta_{ij} = 1$ otherwise. The optimizer's flow plan is recovered and evaluated under the true-cost functional. That is the whole method. Its complexity is one LP solve, about 0.5 s at 200×200 (Table 2), and its surrogate objective is interpretable arc by arc, which stands in contrast to the black-box costings common in swarm-based FCTP work [20], [26], [27].

Two optional polishers refine the proxy plan, both accepting only strict improvements measured under the true cost. The first is an arc-deletion large-neighborhood search [10]: the $k=15$ open arcs carrying the smallest flow are, one at a time, forbidden (their flow fixed to zero), the restricted proxy LP is re-solved, and a move is kept only if the true cost strictly drops; at most 3 passes are made over the candidate list, stopping early when a pass finds nothing. The intuition is that the proxy's amortization can leave a few thin, barely justified arcs open; deleting one occasionally releases its flow onto cheaper structure. The second polisher is basin hopping [10]: 20 restarts, each kicking the incumbent by closing 3 random open arcs simultaneously, then repairing with the same LNS at $k=25$ and at most 5 passes. Algorithm 1 states the base heuristic with the LNS polish; basin hopping wraps lines 5–11 in a random-restart loop.

Algorithm 1: Proxy + LNS

Input : FCTP-QD instance (c, f, a, b, h or p , tiers); $k = 15$; $P = 3$

Output: flow plan x and its true cost C^*

```

1 for all  $(i, j)$  in  $I \times J$ :  $M_{ij} \leftarrow \min(a_i, b_j)$ 
2  $\delta_{ij} \leftarrow r_2$  if QD instance else 1
3  $\hat{c}_{ij} \leftarrow \delta_{ij} \cdot c_{ij} + f_{ij} / M_{ij}$  (amortized unit cost)
4  $x \leftarrow \text{argmin of penalty-aware LP (8) in } \hat{c}$ 
5  $C^* \leftarrow \text{TRUE-COST}(x)$ 
6 for pass  $\leftarrow 1$  to  $P$  do
7   improved  $\leftarrow$  false
8    $A \leftarrow$  the  $k$  smallest-flow open arcs of  $x$ 
9   for each  $(i, j)$  in  $A$  do
10     $x' \leftarrow \text{argmin of LP (8) with } x_{ij} \equiv 0$  (restricted re-solve)
11    if  $\text{TRUE-COST}(x') < C^*$  then
12       $x \leftarrow x'$ ;  $C^* \leftarrow \text{TRUE-COST}(x')$ ; improved  $\leftarrow$  true

```

```

13   if not improved then exit loop
14   return (x, C*)

```

Three design choices in Algorithm 1 deserve comment. Acceptance is always decided under the true-cost functional, never under the surrogate — the LP is a proposal mechanism, the functional is the judge. The destroy set targets small-flow arcs rather than random ones, because thin arcs are where amortization errs most visibly: f_{ij}/M_{ij} understates the effective charge when the realized flow falls far short of M_{ij} . And the pass cap of 3, like the caps in basin hopping, exists because the measured marginal gains decay quickly; §5 quantifies what the polishers add (and, at 200×200 , what they cost in time).

4.6. The ACO Analyzers

The ant colony configuration follows the canonical framework [8], [9], adapted to transportation structure in the spirit of [13], [22], with an embedded local search as in [19]. The parameters are fixed across all experiments: 30 ants per iteration, at most 100 iterations, pheromone exponent $\alpha = 1$, heuristic exponent $\beta = 2$, evaporation rate $\rho = 0.5$, and early stopping after 25 iterations without improvement. Pheromone deposit is elitist and flow-proportional (pheromone τ , distinct from the discount thresholds τ_2, τ_3): the iteration-best ant adds $\tau_{ij} += x_{ij}/Z$ on its arcs, with Z a normalizing constant, so widely used arcs accumulate reinforcement in proportion to the flow they carry. Heuristic information η_{ij} is the amortized unit cost of §4.5 — the swarm and the proxy thus see the same arc prices, which makes later attribution clean. Initial pheromone τ_0 is relaxation-informed, derived from the proxy LP's solution structure rather than set uniformly. Construction uses a gumbel-argmax rule over $\tau^\alpha \eta^{-\beta}$ -weighted arc scores, and each constructed plan passes through a feasible 4-cycle local search — the classical transportation-tableau improvement step [3], [4] — with $20(m+n)$ trials per iteration, every trial delta-evaluated under the full true cost. On QD instances, a tier-snapping rule moves any flow lying within 15% below a tier threshold $\tau_2 M_{ij}$ or $\tau_3 M_{ij}$, snapping it up to the threshold when feasibility allows, since all-unit discounts reward exact threshold crossings.

The decisive design element is dual seeding. Before the colony runs, both the proxy plan and the remedy plan are scored under true cost; the better one is injected twice — into the initial pheromone τ_0 and into the first elite slot, so that the seed plan itself is always among the candidates the elitist update can return. This yields a dominance property by construction: $\text{ACO} \leq \text{best seed}$ on every run. The property converts the swarm from a competitor into an analyzer. Whatever gap appears between its output and its seed is a measured, attributable effect rather than a claim.

A second instrument sharpens the attribution. The seed-hierarchy protocol repeats otherwise identical runs with four different seed sources, spanning the classical constructors: NW, the northwest-corner rule, which fills the transportation tableau from the top-left cell, exhausting supplies and demands in index order [3], [4]; LCM, the least-cost method, which allocates greedily to the cheapest remaining cell; VAM, Vogel's approximation method, which prioritizes the row or column with the largest penalty difference between its two cheapest cells [3], [7]; and LP, the amortized-proxy plan of §4.5. Section 5.4 reads the resulting seed-inheritance anatomy off this protocol.

The unseeded arm, ACO-NS, runs the identical configuration with no plan seed — the relaxation-informed τ_0 is retained — and is the only pipeline arm exhibiting run-to-run variance (NW-seeded ablation runs excepted). One further operator was implemented and evaluated: an arc-consolidation move that merges thin arcs onto surviving structure. It changed the outcome by 0.00 on every tested instance, was dropped from the pipeline, and is reported once in §5 and never again.

4.7. Instances and Evaluation Protocol

Nine base instances span three orders of magnitude of problem size. I1–I3 (3×4 , 5×5 , 8×6) are hand-specified and reproduced verbatim in Appendix A. I4–I9 are generated with NumPy RandomState under fixed seeds (789, 1011, 1213, 1415, 1617, 1819) at sizes 10×10 , 20×20 , 30×30 , 50×50 , 100×100 , and 200×200 ; the base side draws integer quantities from $U(80, \dots, 120)$, the other side from $U(70, \dots, 110)$ normalized so its total equals $1 \pm r$ times the base total, with imbalance direction and ratio $r \in \{0.15, 0.20\}$ as tabulated in Appendix A. Unit costs are integer uniforms over instance-specific ranges from 1,30 to 1,50; penalties are $h_i \sim U\{2, \dots, 15\}$ or $p_j \sim U\{10, \dots, 50\}$ by direction. Fixed charges come in two regimes: nominal, $f_{ij} \sim U\{30, \dots, 120\}$, and realistic (bigF), scaled tenfold to $U(300, \dots, 1200)$ with seed offset +5000, giving the variants I4_bigF through I9_bigF. Quantity-discount variants I2_QD, I7_QD, and I9_QD apply the all-unit tier structure of §3 — thresholds $\tau_2 = 0.25$ and $\tau_3 = 0.5$ of M_{ij} , rates $r_2 = 0.85$ and $r_3 = 0.70$ — under nominal fixed charges. The 200×200 frontier is thus probed in three forms: nominal-F I9, realistic-F I9_bigF, and I9_QD.

Run counts reflect determinism. The six deterministic pipelines (Dummy, Remedy, MILP, Proxy, Proxy+LNS, Proxy+BH) are run once per instance — repetition would reproduce the number, not estimate it. ACO and ACO-NS are stochastic and run 10 times per configuration, reduced to 3 runs at I8 and 2 at I9_QD where per-run cost grows; the seed-hierarchy protocol uses 5 runs per seed source. Statistical comparisons of the unseeded arm against the seeded constant use the one-sample Wilcoxon signed-rank test with the one-sided alternative “greater” (unseeded cost exceeds seeded cost), reported with the exact p-value, Cohen's d effect size, and n ; configurations with $n \leq 3$ are marked descriptive. Where the seeded arm is deterministic, its standard deviation is exactly zero; we report this variance collapse as a finding in its own right rather than papering over it, and such comparisons are stated as exact ties. All timings are wall-clock seconds on a single workstation; we report solver, budgets, seeds, and run counts explicitly, and we treat the 120 s MILP budget — not any hardware constant — as the reproducibility-relevant quantity.

5. Computational Results

5.1. Experimental Overview

The benchmark set comprises nine instances in three groups. The first group, I1–I3 (3×4 , 5×5 , 8×6), consists of small hand-set instances whose full cost, supply, demand, penalty, and fixed-charge matrices are reproduced verbatim in Appendix A; every pipeline solves them in under one second, and they function mainly as correctness checks for the true-

cost evaluator. The second group, I4–I8 (10×10 to 100×100), is generated under a seeded protocol — costs $c_{ij} \sim U\{lo, hi\}$, base-side quantities drawn as integers from $U80,120$, the opposite side from $U70,110$ and normalized to $(1 + \text{ratio}) \times$ the base total, holding costs $h_i \sim U\{2,15\}$, shortage penalties $p_j \sim U\{10,50\}$ — with the imbalance direction alternating between excess demand (ED) and excess supply (ES): I4 (10×10 , seed 789, ED 0.20), I5 (20×20 , seed 1011, ES 0.15), I6 (30×30 , seed 1213, ED 0.15), I7 (50×50 , seed 1415, ES 0.15), and I8 (100×100 , seed 1617, ED 0.20). The third group is the frontier instance I9 (200×200 , seed 1819, ED 0.15). Each instance is evaluated under three cost structures: nominal fixed charges $f_{ij} \sim U\{30,120\}$; realistic fixed charges $f_{ij} \sim U\{300,1200\}$ (denoted bigF); and all-unit quantity discounts (QD) with tier thresholds $\tau_2 = 0.25, \tau_3 = 0.5$ of $M_{ij} = \min(a_i, b_j)$ and tier rates $r_2 = 0.85, r_3 = 0.70$.

A single note covers run counts and timing. The MILP benchmark (HiGHS branch-and-bound) is run once per instance under a 120 s solver budget and returns a certified dual bound; Dummy, Remedy, and Proxy are deterministic LPs and are likewise run once, with proxy solve times of roughly 0.5 s even at 200×200 . The dual-seeded (LP-seeded) ACO is deterministic on every tested instance (standard deviation 0), a variance collapse reported as a finding rather than a convenience; NW-seeded runs retain variance (Table 5), as does the unseeded ACO-NS arm, which is therefore repeated 10 times per configuration (3 runs on I8, 2 on I9_QD). Statistical comparisons use the one-sided Wilcoxon signed-rank test against the seeded constant, with Cohen's d as effect size; configurations with fewer than five runs are reported descriptively.

5.2. The Convention Premium

We first measure what the two linear-era conventions actually cost once the true fixed charges and discount tiers are applied. Table 3 reports true costs for the Dummy, Remedy, and Proxy pipelines against the MILP incumbent on I1–I9 under both fixed-charge regimes; Figure 1 visualizes the resulting premiums.

Table 3. True costs of the Dummy, Remedy, and Proxy pipelines against the MILP incumbent on I1–I9 under nominal fixed charges ($f_{ij} \sim U\{30,120\}$) and realistic fixed charges ($f_{ij} \sim U\{300,1200\}$).

Instance	Dummy	Remedy	Proxy	MILP incumbent
I1, nominal F	6,234.00	5,931.00	5,931.00	5,931.00 (exact)
I2, nominal F	10,833.56	9,162.88	9,162.88	9,162.88 (exact)
I3, nominal F	14,497.31	12,220.60	12,220.60	12,209.79 (exact)
I4, nominal F	15,332.69	11,518.60	11,457.16	11,457.16 (exact)
I5, nominal F	10,036.72	9,530.11	9,448.58	9,077.46 (exact)
I6, nominal F	30,166.54	21,475.12	21,258.95	20,936.03 (0.43% gap)
I7, nominal F	22,288.01	19,617.34	19,401.12	18,423.90 (0.97% gap)
I8, nominal F	84,033.13	53,696.72	53,084.30	44,458.76 (3.50% gap)
I9, nominal F	140,063.45	85,414.49	82,949.23	82,946.60 (8.67% gap)
I4, realistic F	27,716.69	23,245.60	22,811.92	20,327.93 (exact)
I5, realistic F	33,832.72	35,000.11	31,701.72	24,413.82 (exact)
I6, realistic F	66,868.54	58,726.12	51,978.99	42,782.91 (1.07% gap)
I7, realistic F	78,601.01	78,549.34	64,899.68	48,981.06 (exact)
I8, realistic F	198,486.13	176,231.72	137,952.73	124,202.81 (15.73% gap)
I9, realistic F	378,626.45	327,892.49	234,653.92	234,633.10 (23.77% gap)

Read against the MILP incumbent column, the uniform-dummy convention carries a premium of +5.11% already at 3×4 , and the premium compounds with scale and charge magnitude, reaching +89.01% at I8 and +68.86% at I9 under nominal charges. The incumbent column makes these premiums directly verifiable: every Dummy, Remedy, and Proxy figure can be checked against the certified reference on its own row. A closer look reveals a structural pattern the aggregate range conceals: the excess-demand instances pay far more than the excess-supply ones. Under nominal F the ED instances I4, I6, I8, I9 carry premiums of +33.83%, +44.09%, +89.01%, +68.86%, while the ES instances I5 and I7 pay only +10.57% and +20.97%. Because shortage penalties ($p_j \sim U\{10,50\}$) are roughly triple the holding costs ($h_i \sim U\{2,15\}$), a convention that prices all imbalance through one mean-penalty dummy systematically misprices the expensive side, and the error grows with the number of arcs the dummy plan opens. Under realistic charges the asymmetry washes out — the dummy premium sits between +36.35% and +61.37% regardless of imbalance direction — because the fixed-charge component, which the dummy LP ignores entirely, comes to dominate the objective. The Remedy behaves better, as expected, yet its advantage erodes exactly where it matters: its gap to the incumbent grows from +0.54% (I4) to +20.78% (I8) under nominal F, and from +14.35% to +60.37% under realistic F. On I5_bigF the Remedy is even worse than the Dummy (35,000.11 against 33,832.72), a reversal worth noting: penalty awareness without fixed-charge awareness is not a strict improvement once F is large.

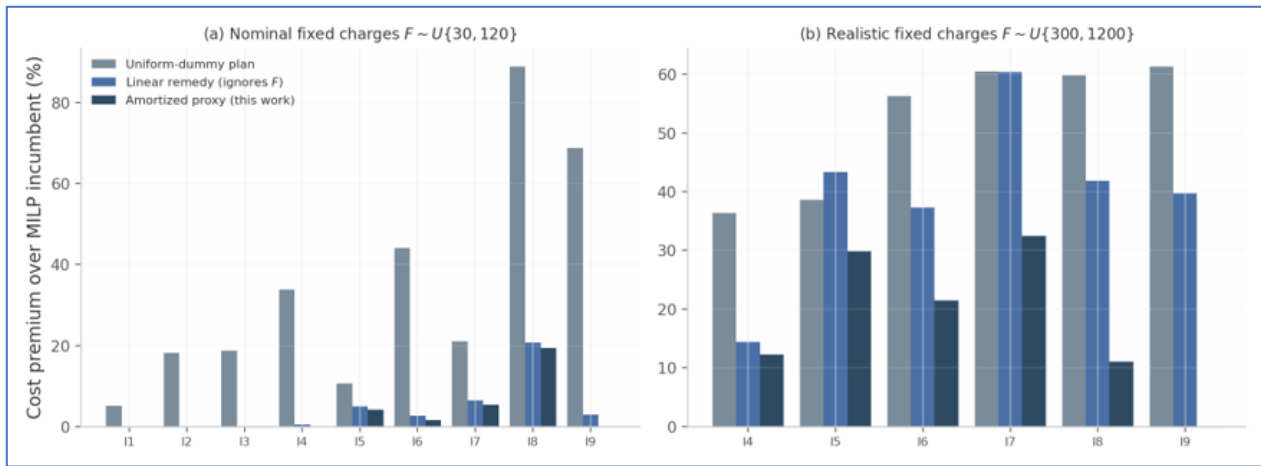


Figure 1: Convention premiums over the MILP incumbent. Left panel: Dummy and Remedy premiums under nominal fixed charges; right panel: premiums under realistic fixed charges.

Figure 1 makes the compounding visible. The left panel shows two diverging curves: the Remedy premium climbs slowly and irregularly, whereas the Dummy premium accelerates after 30×30 and more than quadruples between 50×50 and 100×100 . The right panel widens the gap. With fixed charges scaled by ten, both linear conventions detach from the incumbent frontier, and their curves largely overlap — penalty structure stops discriminating between them because both ignore the term that now dominates the cost. The Proxy curve, included in both panels for reference, stays close to zero except at I8; we return to that outlier in Section 5.5. The visual summary matches the numeric one: the convention premium is not a fixed tax but a quantity that compounds with charge magnitude and problem scale, from a tolerable 5.11% on a 3×4 textbook instance to a prohibitive 89.01% at operational scale.

5.3. Measured Limits of Exact Solving

The FCTP is NP-hard since Balinski’s formulation [24], so exact solving must eventually degrade; the question is where, and by how much, on instances of practical size. Table 4 summarizes the certified optimality gap of the HiGHS branch-and-bound after the 120 s budget across the three cost structures, and Figure 2 plots the same trajectory.

Table 4. MILP certified gap after the 120 s solver budget (“exact” denotes proven optimality, with solve time).

Size	Nominal F	Realistic F	QD
5×5	—	—	exact (100 binaries)
10×10	exact, 1.5 s	exact, 1.73 s	—
20×20	exact, 3.3 s	exact, 8.0 s	—
30×30	0.43%	1.07%	—
50×50	0.97%	exact, 56.6 s	3.32% (10,000 binaries)
100×100	3.50%	15.73%	—
200×200	8.67%	23.77%	no feasible incumbent (160,000 binaries)

Three regimes emerge. Up to 20×20 the solver certifies optimality in seconds, and the exact model is simply the right tool. From 30×30 to 100×100 the certified gap widens roughly with the number of binary variables — 0.43% to 3.50% under nominal F — yet incumbents remain strong and the MILP still dominates every heuristic we tested at mid scale. At 200×200 the picture changes qualitatively. The nominal-F run closes only to an 8.67% gap (incumbent 82,946.60, dual 75,754.24), the realistic-F run to 23.77% (incumbent 234,633.10, dual 178,856.46), and the QD model returns no feasible incumbent within the 120 s solver budget (≈ 140 s wall including ≈ 19 s model build; 160,000 binaries). We stress the limits of this observation: it is a measurement of one open-source solver under one budget, not a claim about commercial solvers, which we did not test.

Against this ceiling the Proxy’s frontier numbers deserve careful phrasing. At 200×200 the proxy LP costs 82,949.23 under nominal F and 234,653.92 under realistic F — that is +0.003% and +0.01% above the respective 120 s MILP incumbents, obtained in about 0.5 s instead of the full 120 s budget. We read these as numerically indistinguishable from the 120 s MILP incumbent, at ~ 0.5 s, and nothing more: with dual bounds 8.67–23.8% away, neither pipeline is certified near-optimal at this scale, and every frontier statement in this paper compares incumbents, not optima. The QD frontier is starker. With the MILP returning no feasible incumbent within the 120 s solver budget, the proxy’s 76,859.52 (LP in seconds; a full seeded ACO run takes 25.7 s) is the only competitive plan produced by any pipeline we tested at 200×200 QD: the next-best linear pipeline, the Remedy, reaches 79,839.69, the Dummy 134,611.33, and unseeded ACO delivers 145,009.08 on average (best run 143,217.91).

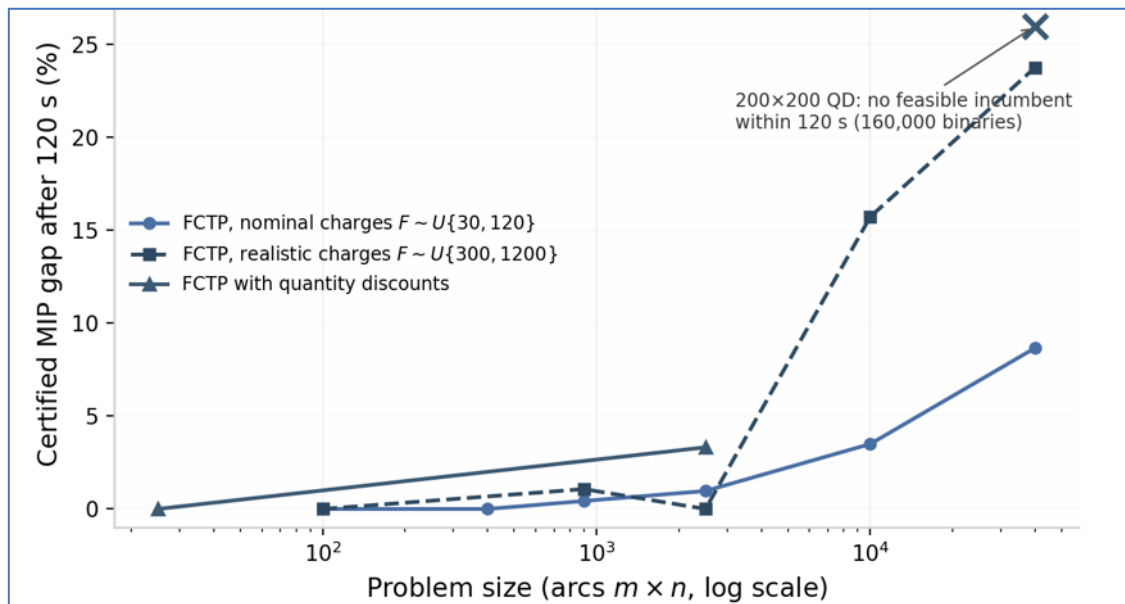


Figure 2: Certified MILP optimality gap after 120 s as a function of problem size, for nominal fixed charges, realistic fixed charges, and quantity discounts.

Figure 2 condenses the scalability measurement into three curves. All three sit at zero through 20×20, then bend upward at rates that track the difficulty of the cost structure: nominal F rises gently to 8.67%, realistic F steepens after 50×50 and ends at 23.77%, and the QD curve terminates at 200×200 not in a large gap but in the absence of any incumbent at all. The figure replaces a piece of folklore. The often-repeated claim that “exact methods fail” on fixed-charge transportation is, on our measurements, wrong below 100×100 and imprecise above it: what actually happens is a monotonic erosion of the certified bound, followed — only for the non-convex QD model at the largest scale — by the solver’s failure to find even a feasible plan within the budget. That distinction matters for practice, because it identifies 100×100–200×200 as the band where heuristic incumbents stop being conveniences and become the only available answers.

5.4. The Seed-Inheritance Anatomy of ACO

The results above already contain an anomaly that deserves its own experiment. Table 3’s Proxy column hints at it, and Table 5’s LP row below records it directly: the dual-seeded ACO — 30 ants, up to 100 iterations, $\alpha = 1$, $\beta = 2$, $\rho = 0.5$, elitist flow-proportional deposits, amortized heuristic information, relaxation-informed initial pheromone, feasible 4-cycle local search with $20(m + n)$ trials per iteration — returned, on every tested instance, exactly its LP seed (± 0.00), raw LP seed cost coinciding with the post-swarm result on all five configurations. The swarm’s marginal contribution beyond the seed was precisely zero. To find out whether this is a property of the LP seed specifically or of the algorithm generally, we ran a seed-hierarchy ablation: four seed sources — northwest corner (NW), least-cost method (LCM), Vogel’s approximation (VAM), and the amortized LP — each used to initialize five runs of the identical ACO on five configurations (I6, I7, I6_bigF, I7_bigF, I7_QD). Construction heuristics of this kind are the standard initializers in the transportation-ACO literature [13], [14], [22], and VAM in particular is often credited as a strong starting point for both exact and metaheuristic pipelines [7], [17]. Table 5 reports raw seed costs and post-swarm results; Figure 3 visualizes the trajectories.

Table 5. Seed-hierarchy ablation. Each cell gives the raw seed’s true cost and the seeded ACO result (mean $\pm$ std over 5 runs).

Config	NW raw $\rightarrow$ ACO	LCM raw $\rightarrow$ ACO	VAM raw $\rightarrow$ ACO	LP raw $\rightarrow$ ACO
I6	77,020.28 $\rightarrow$ 30,908.94 $\pm$ 432.24	31,493.70 $\rightarrow$ 31,376.15 $\pm$ 262.84	32,349.50 $\rightarrow$ 31,035.88 $\pm$ 255.25	21,258.95 $\rightarrow$ 21,258.95 $\pm$ 0
I7	140,034.71 $\rightarrow$ 29,035.10 $\pm$ 1,097.54	25,032.56 $\rightarrow$ 25,032.56 $\pm$ 0	26,685.97 $\rightarrow$ 26,685.97 $\pm$ 0	19,401.12 $\rightarrow$ 19,401.12 $\pm$ 0
I6_bigF	114,829.28 $\rightarrow$ 62,860.00 $\pm$ 1,762.83	54,822.35 $\rightarrow$ 54,822.35 $\pm$ 0	64,613.67 $\rightarrow$ 62,868.33 $\pm$ 1,046.61	51,978.99 $\rightarrow$ 51,978.99 $\pm$ 0
I7_bigF	203,844.71 $\rightarrow$ 83,910.54 $\pm$ 2,008.10	70,248.14 $\rightarrow$ 70,248.14 $\pm$ 0	73,058.65 $\rightarrow$ 73,058.65 $\pm$ 0	64,899.68 $\rightarrow$ 64,899.68 $\pm$ 0
I7_QD	107,583.62 $\rightarrow$ 26,173.29 $\pm$ 870.51	22,070.32 $\rightarrow$ 22,070.32 $\pm$ 0	23,995.14 $\rightarrow$ 23,995.14 $\pm$ 0	16,857.99 $\rightarrow$ 16,857.99 $\pm$ 0

Three patterns stand out. First, the strong seeds are absorbing states. The LP seed passes through the swarm unchanged (± 0) on all five configurations; the LCM seed does so on four of five (it moves only on I6) and the VAM seed on three of five (it moves on I6 and I6_bigF). Each of the three exceptions improves its seed by only 0.37–4.06%, and none reaches the LP basin. These seeds behave as 4-cycle local optima — measured directly for the proxy plans in §5.5 and evidenced for LCM and VAM by their pass-through behavior — and the swarm, whose improvement machinery is exactly that 4-

cycle local search plus pheromone-biased reconstruction, has nothing to add to such a plan. Second, the weak NW seed tells the symmetric story. The swarm recovers between 45.3% and 79.3% of the NW seed's cost (for instance, 77,020.28 → 30,908.94 on I6), a genuinely large improvement; yet every NW-seeded run stalls in the cost band of the LCM/VAM seeds — on I6 it even lands marginally below the LCM-seeded result — and no run, on any configuration, approaches the LP basin. Third, final quality is inherited, not created: the LP seed's first-place rank is preserved on all five configurations, and the only reordering in the entire table is a single adjacent swap between the two mid-tier seeds on I6 (VAM final 31,035.88 against LCM final 31,376.15). The ranking of outputs is, to within one adjacent pair, the ranking of inputs.

The unseeded arm completes the anatomy. Without LP seeds (relaxation-informed initial pheromone retained), the same ACO is +9.3 to +88.7 worse than the seeded constant across the ten measured configurations — for example, 84,237.40 against the proxy's 64,899.68 on I7_bigF. The difference is significant at Wilcoxon signed-rank $p = 9.77 \times 10^{-4}$ ($n = 10$, one – sided) on the eight configurations with full run counts, with Cohen's d between 6.7 and 91.3 — effect sizes so large that the test is almost a formality — and remains adverse descriptively where runs are few (I8: +72.1, $p = 0.125$, $n = 3$; I9_D: +88.7, $p = 0.25$, $n = 2$). Dual-seeded (LP-seeded) runs, recall, have standard deviation 0: the variance collapse in that arm is itself the signature of an algorithm whose search never leaves its initialization. The cause is visible in the construction step. Each ant builds a plan with up to $m+n-1$ open arcs, and the 4-cycle local search rarely empties an arc once opened; under realistic fixed charges, where the F-component dominates the objective, the unseeded swarm's excess cost over the proxy is well approximated by the number of unnecessarily open arcs times the mean fixed charge. An arc-consolidation operator designed to close such arcs showed a 0.00% effect on every instance and was dropped from the pipeline.

We state the interpretive claim analytically, with its boundaries attached. Within the tested budget, on these instances, for this configuration family, the swarm adds no quality beyond its initialization. Whatever the seeded ACO returns was already present in its seed, and what the literature reports as swarm performance on transportation problems [8], [9], [13], [14], [22] may, when initialized with VAM- or NW-type constructions and evaluated without seed ablation, have measured seed basins rather than swarm capability. This is not a claim that ACO cannot help anywhere — the no-free-lunch theorem forbids that generality [15], and our own future-work agenda identifies dynamic multi-period variants where swarm methods may genuinely matter. It is a claim that the contribution of the swarm component is measurable, that we measured it at 0.00%, and that any future report of swarm superiority in this problem class should be accompanied by the same ablation.

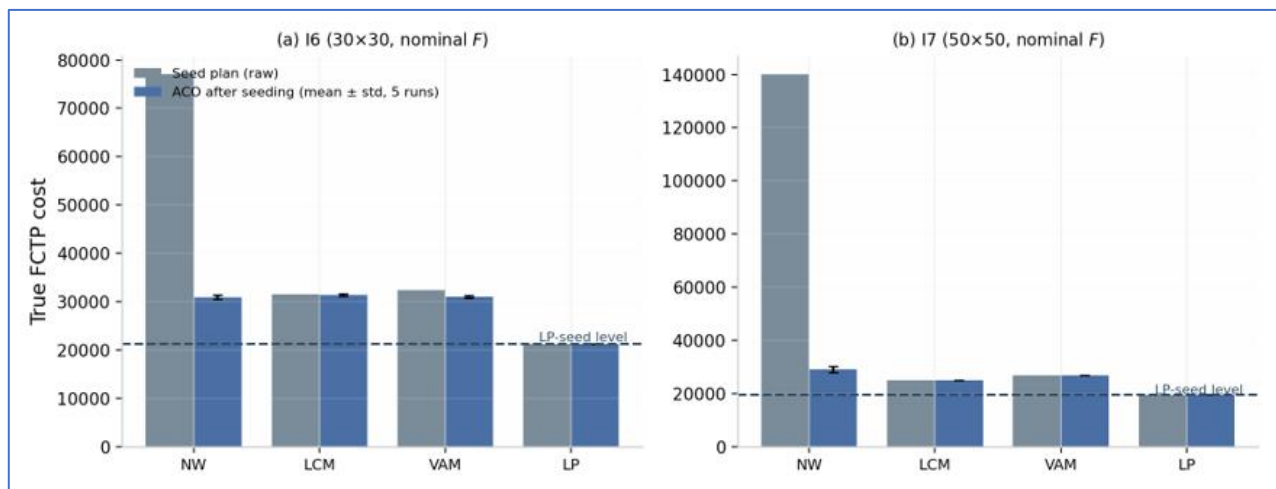


Figure 3: Seed inheritance in ACO. Raw seed true costs (markers) and post-swarm results (bars) for NW, LCM, VAM, and LP seeds across five configurations; the LP seed's basin is unreachable from below.

Figure 3 displays the mechanism in one view. The NW trajectories plunge impressively — a casual reading might celebrate the swarm's 45–79% recovery — but every trajectory asymptotes at the heuristic band, far above the flat LP line. The LCM and VAM lines are nearly flat as well, and the LP line is exactly flat, with zero-width error bars that distinguish it visually from every other series. What the figure makes obvious, and what a results table without seed ablation would hide, is that the four output levels were determined before the first ant moved. The swarm's iteration count, its pheromone parameters, its elitism — none of these leave a trace on the output once the seed is fixed. That is the sense in which we describe ACO here as an analyzed baseline whose anatomy, not its performance, is the contribution.

5.5. Polish and Near-Optimality

The remaining question is whether the proxy plan can be improved cheaply, and how close the full pipeline sits to the exact frontier where that frontier is still visible. Table 6 reports the two polish operators — arc-deletion large-neighborhood search (LNS) and basin hopping (BH) — on the five configurations where they were run.

Table 6. Polish results. True costs of Proxy, Proxy+LNS ($k = 15\text{smallest} - \text{flowarcs}, \leq 3\text{passes}$), and Proxy+BH ($20\text{kicksof}3\text{randomarcclosures} + \text{LNSpolish}$).

Config	Proxy	Proxy+LNS	Proxy+BH
I4 bigF	22,811.92	22,577.24	22,577.24
I5 bigF	31,701.72	30,994.85	30,452.14
I6 bigF	51,978.99	51,430.53	50,779.63
I7 bigF	64,899.68	64,544.41	63,850.22
I7 QD	16,857.99	16,839.33	16,755.69

The polish is real but modest: LNS removes between 0.1% (I7_QD) and 2.2% (I5_bigF) of the proxy cost, and basin hopping extends the total gain to at most 3.9% (I5_bigF: 31,701.72 $\rightarrow$ 30,452.14). These are the gains of a local method working on an already local-optimal plan, and their smallness is itself informative. A direct stability measurement confirms the interpretation: subjecting proxy plans to $20(m + n)$ random 4-cycle trials, repeated five times, yields zero improving moves on I5, I7_bigF, I9_bigF, I7_QD, and I9_QD, and a single improving move of -0.087% on I8. The proxy plan is, by measurement, a 4-cycle local optimum — which also explains, retrospectively, why the seeded swarm of Section 5.4 finds nothing to improve: its search operator and the proxy's stability set coincide. Where the proxy is far from the incumbent — I8 under both regimes, with gaps of $+19.40\%$ and $+11.07\%$, and I7_bigF at $+32.50\%$ — the polish recovers only part of the distance, marking these as the configurations where better improvement operators (ejection chains are a candidate) are genuinely needed rather than decorative.

The near-optimality statement belongs to the frontier. On I9 under nominal F, the four pipelines span a wide band: the Dummy costs 140,063.45 ($+68.86\%$), the Remedy 85,414.49 ($+2.98\%$), the MILP incumbent after 120 s is 82,946.60 at a certified gap of 8.67%, and the Proxy costs 82,949.23 — 0.003% above a 120 s incumbent, delivered in about 0.5 s. Under realistic F the corresponding tie is 234,653.92 against 234,633.10 ($+0.01\%$). We stop short of calling the proxy near-optimal here: with the dual bound 8.67% below the incumbent, optimality at 200×200 is unknown, and what we report is an incumbent tie at a fraction of the computational budget. Within the tested budget and on these instances, that is the strongest statement the evidence supports — and, given the 23.77% certified gap under realistic charges, it is also a statement about how little any method currently knows at this scale.

6. Discussion

6.1. Interpretation of Results

Taken together, the measurements shift explanatory weight away from algorithms and toward modeling decisions. The uniform-dummy convention — a bookkeeping device inherited from the balanced textbook problem [3] — costs $+5.11\%$ on the smallest instance and $+89.01\%$ at 100×100 under nominal fixed charges, and never less than $+36.35\%$ under realistic ones (Tables 3 and 4). The penalty-aware remedy holds wherever charges are mild, with gaps of 0–6.48% through 50×50 under nominal F, yet it breaks down as charges grow: $+14.35\%$ at 10×10 rising to $+60.37\%$ at 50×50 under realistic F. The pattern has a structural reading. Fixed charges reward consolidation, and both the dummy plan and the remedy plan spread flow across many arcs because neither objective sees f_{ij} . Each unnecessarily open arc then pays a full charge, and the excess scales with arc count. The proxy survives precisely because its amortization internalizes that driver: pricing each arc at $c_{ij} + f_{ij}/M_{ij}$ lets the LP see the charge before opening the arc, rewarding consolidation inside the optimization, while the remedy, blind to f_{ij} , cannot make that trade at any scale. A convention choice thus behaves as a first-order cost driver — of the same order as the choice of optimizer itself.

At the frontier, the amortized proxy reorganizes the exact-heuristic boundary. On I9 it returns plans numerically indistinguishable from the 120 s MILP incumbent, at ~ 0.5 s — $+0.003\%$ under nominal F, $+0.01\%$ under realistic F, roughly $1/240$ of the solver budget. On I9 with quantity discounts it is the only pipeline returning a competitive plan: 76,859.52 in seconds, against 79,839.69 for the next-best linear pipeline, while the MILP returns no feasible incumbent within the 120 s solver budget (160,000 binaries). One caveat keeps this claim precise. The proxy ties an incumbent, not an optimum; dual bounds at 200×200 sit 8.67–23.8% below the incumbents, so nothing at that scale is certified near-optimal. Still, the practical comparison at 200×200 is no longer exact-versus-heuristic. It is budget-constrained incumbent versus a seconds-fast approximation of it.

The seed-inheritance anatomy, finally, changes how swarm results in this space should be read. On every tested instance the dual-seeded ACO returns its LP seed unchanged (± 0.00), while the unseeded swarm runs $+9.3\%$ to $+88.7\%$ worse ($Wilcoxonp = 9.77 \times 10^{-4}$, $n = 10$; §5.4). Prior ACO studies on transportation problems report final swarm quality with heuristic initialization folded in [9], [22], [26]; the initialization and the swarm's own contribution have not been separated. Our protocol separates them, and the separation is unambiguous: final quality rank equals seed rank, LCM and VAM seeds pass through the swarm essentially untouched, Northwest-corner recovery (-60%) stalls at the heuristic basin, and dual-seeded (LP-seeded) runs are deterministic (std 0). This does not show prior results to be wrong — the no-free-lunch theorems preclude any such blanket verdict [15] — but it shows that an apparently swarm-produced advantage may be entirely inherited. Reporting practice should follow.

6.2. Limitations

Several boundaries delimit these claims. First, every MILP number comes from one engine, HiGHS, under a 120 s budget — the 200×200 QD run spent roughly 19 s building the 160,000-binary model (build time is never folded into the 120 s solver budget) and about 140 s of wall time overall without finding a feasible incumbent. Commercial solvers may push the exact frontier outward; the certified gaps and the 200×200 silence are engine- and budget-specific statements, not

properties of the problem class. Second, the instances are synthetic: I1–I3 are hand-specified and I4–I9 are generated under fixed seeds with uniform cost, supply, and penalty ranges, and only two fixed-charge regimes are tested, $F \sim U(30,120)$ and its tenfold scaling. Real networks exhibit correlated costs and lumpy demands, and calibration against operational data remains future work. Third, the Wilcoxon tests rest on $n=10$ runs per configuration across the eight configurations with full run counts — small samples, however large the measured effect sizes ($d = 6.7 - 91.3$) — and the $I8(n = 3, p = 0.125)$ and $I9_QD(n = 2, p = 0.25)$ entries are descriptive only, licensing no inference. Fourth, frontier claims compare 120 s incumbents, not optima; the associated dual bounds lie 8.67–23.8% away, so statements about the proxy tying the MILP at 200×200 concern equally budget-constrained objects.

Three further limits concern design choices. The discount structure is fixed at two thresholds, $\tau_2 = 0.25$ and $\tau_3 = 0.5$ of M_{ij} , with rates $r_2 = 0.85$ and $r_3 = 0.70$ in the all-unit form of [28]; other schedules will shift the numbers, though not the methodology. The proxy's amortization f_{ij}/M_{ij} is likewise a choice, and its failure mode is visible in our own tables: under extreme fixed-charge dispersion the proxy gap widens to +19.40% (I8) and +32.50% (I7_bigF), and the LNS and basin-hopping polish recover only part of it. Finally, seed inheritance is measured for one ACO configuration family — elitist deposits, 30 ants, 4-cycle local search, $\alpha = 1$, $\beta = 2$, $\rho = 0.5$. Rank-based deposit rules, candidate lists, or deeper local-search neighborhoods may behave differently, and the claim is bounded accordingly.

6.3. Implications and Future Work

Three practical recommendations follow. Retire the uniform dummy: a penalty-aware linear remedy costs nothing extra to solve and removes most of the convention premium wherever fixed charges are mild. In the mid regime, up to 50×50 , the budgeted MILP remains the default — it certifies gaps under 1% there under nominal F — and the proxy serves as its cheap first incumbent. At 100×100 and beyond, run the amortized proxy — with LNS or basin-hopping polish where seconds-to-minutes are available — rather than waiting on incumbents that may never arrive. And swarm studies on transportation problems should report seed ablations as standard; without them, the source of any measured advantage is unknowable.

On the research side, the four directions of §7.2 follow from these measurements; commercial-solver comparisons and incremental (rather than all-unit) discount forms would further sharpen the frontier measurements reported here.

7. Conclusion

7.1. Summary of Contributions

This study set out to replace two pieces of transportation folklore — that exact methods fail under fixed charges and quantity discounts, and that swarm intelligence succeeds there — with certified measurement. Five findings emerged. First, the uniform-dummy convention, harmless folklore in the linear case, exacts a premium of 5.11% to 89.01% under fixed charges, and the heterogeneous linear remedy breaks down by +14.35% to +60.37% once realistic charge magnitudes arrive; linear modeling habits do not survive nonlinearity. Second, a modern MILP solver has a measurable ceiling: certified gaps widen from 0% to 23.77% across our frontier, and at 200×200 with quantity discounts the solver returns no feasible incumbent within the 120 s budget (≈ 140 s wall including ≈ 19 s of model build; 160,000 binaries) — a concrete, reproducible statement where vague failure claims once stood.

Third, the amortized fixed-charge proxy LP closes much of that gap in seconds. It is numerically indistinguishable from the 120 s MILP incumbent at 200×200 (+0.003% nominal, +0.01% realistic) at roughly 0.5 s, and under quantity discounts at 200×200 it is the only pipeline returning a competitive plan (76,859.52, against 79,839.69 for the next-best linear pipeline). Fourth, the seeded ACO anatomy is unambiguous: the swarm adds exactly 0.00% beyond its LP seed on every instance, final rank equals seed rank, dual-seeded (LP-seeded) runs are deterministic (std 0), and unseeded runs proliferate arcs. The swarm's quality is the seed's quality. Fifth, the protocol itself — every heuristic compared against a certified dual bound, every swarm claim ablated against its seed — is portable, and it adds seconds, not minutes.

7.2. Future Work

The natural extension is the dynamic, multi-period FCTP with inventory carryover and time-varying penalties, where MILP dimensionality genuinely explodes and swarm methods may finally matter — provided they face the same certified-benchmark discipline imposed here. Three further directions follow: LP-seeded adaptive pheromone mechanisms, which test whether seed inheritance survives a smarter update rule; hybrid exact-heuristic frontier methods, coupling budgeted MILP incumbents with amortized-proxy LPs and stronger operators such as ejection chains; and multi-echelon extensions of the unbalanced FCTP-QD, which would carry the same one-functional protocol into richer network structures.

The fixed-charge transportation problem has been NP-hard since Balinski's formulation [24], and no budget of engineering will change that. What measurement can change is how accurately we report the frontier — and, at 200×200 , a proxy LP already answers in half a second.

References

- [1] F. L. Hitchcock, "The distribution of a product from several sources to numerous localities," *Journal of Mathematics and Physics*, vol. 20, no. 1–4, pp. 224–230, 1941.
- [2] T. C. Koopmans, "Optimum utilization of the transportation system," *Econometrica*, vol. 17, pp. 136–146, 1949.
- [3] F. S. Hillier and G. J. Lieberman, *Introduction to Operations Research*, 11th ed. New York, NY, USA: McGraw-Hill, 2021.
- [4] H. A. Taha, *Operations Research: An Introduction*, 10th ed. Harlow, UK: Pearson, 2017.
- [5] M. E. B. Brigden, "A variant of the transportation problem in which the constraints are of mixed type," *Operational Research Quarterly*, vol. 25, no. 3, pp. 437–445, 1974.
- [6] P. Pandian and G. Natarajan, "A new approach for solving transportation problems with mixed constraints," *Journal of Physical Sciences*, vol. 14, pp. 53–61, 2010.

- [7] M. M. Ahmed, N. Sultana, A. R. Khan, and S. Uddin, "An innovative approach to obtain an initial basic feasible solution for the transportation problems," *Journal of Physical Sciences*, vol. 22, pp. 23–42, 2017.
- [8] M. Dorigo, *Optimization, Learning and Natural Algorithms*, Ph.D. dissertation, Politecnico di Milano, Milan, Italy, 1992.
- [9] M. Dorigo and T. Stützle, "Ant colony optimization: Overview and recent advances," in *Handbook of Metaheuristics*, F. Glover and G. Kochenberger, Eds. New York, NY, USA: Springer, 2019, pp. 311–351.
- [10] C. Blum and A. Roli, "Metaheuristics in combinatorial optimization: Overview and conceptual comparison," *ACM Computing Surveys*, vol. 35, no. 3, pp. 268–308, 2003.
- [11] J. H. Holland, *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI, USA: University of Michigan Press, 1975.
- [12] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. IEEE Int. Conf. Neural Networks*, Perth, WA, Australia, 1995, pp. 1942–1948.
- [13] R. Jovanović, M. Tuba, and S. Voß, "An ant colony optimization algorithm for partitioning graphs with supply and demand," *Applied Soft Computing*, vol. 41, pp. 317–330, 2016.
- [14] D. Teodorović, "Swarm intelligence systems for transportation engineering," *Transportation Research Part C: Emerging Technologies*, vol. 16, no. 6, pp. 651–667, 2008.
- [15] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, 1997.
- [16] T. Stamadianos, A. Taxidou, M. Marinaki, and Y. Marinakis, "Swarm intelligence and nature inspired algorithms for solving vehicle routing problems: A survey," *Operational Research*, vol. 24, no. 3, Art. no. 47, 2024.
- [17] F. A. Wireko, I. D. K. Mensah, E. N. A. Aborhey, S. A. Appiah, C. Sebil, and J. Ackora-Prah, "The maximum range method for finding initial basic feasible solution for transportation problems," *Results in Control and Optimization*, vol. 19, Art. no. 100551, 2025.
- [18] M. Korzeń and I. Gisterek, "Applying ant colony optimization to reduce tram journey times," *Sensors*, vol. 24, no. 19, Art. no. 6226, 2024.
- [19] H. Wu and Y. Gao, "An ant colony optimization based on local search for the vehicle routing problem with simultaneous pickup–delivery and time window," *Applied Soft Computing*, vol. 139, Art. no. 110203, 2023.
- [20] M. B. Fakhrzad, F. Goodarzian, and A. M. Golmohammadi, "Addressing a fixed charge transportation problem with multi-route and different capacities by novel hybrid meta-heuristics," *Journal of Industrial and Systems Engineering*, vol. 12, no. 1, pp. 167–184, 2019.
- [21] R. Elshaer and H. Awad, "A taxonomic review of metaheuristic algorithms for solving the vehicle routing problem and its variants," *Computers & Industrial Engineering*, vol. 140, Art. no. 106242, 2020.
- [22] E. M. U. S. B. Ekanayake, S. P. C. Perera, W. B. Daundasekara, and Z. A. M. S. Juman, "A modified ant colony optimization algorithm for solving a transportation problem," *Journal of Advances in Mathematics and Computer Science*, vol. 35, no. 5, pp. 83–101, 2020.
- [23] K. Govindan, A. Jafarian, and V. Nourbakhsh, "Designing a sustainable supply chain network integrated with vehicle routing: A comparison of hybrid swarm intelligence metaheuristics," *Computers & Operations Research*, vol. 110, pp. 220–235, 2019.
- [24] M. L. Balinski, "Fixed-cost transportation problems," *Naval Research Logistics Quarterly*, vol. 8, no. 1, pp. 41–54, 1961, doi: 10.1002/nav.3800080104.
- [25] V. Adlakha and K. Kowalski, "A simple heuristic for solving small fixed-charge transportation problems," *Omega*, vol. 31, no. 3, pp. 205–211, 2003, doi: 10.1016/S0305-0483(03)00025-2.
- [26] M. Hajiaghayi-Keshteli, M. Molla-Alizadeh-Zavardehi, and R. Tavakkoli-Moghaddam, "Addressing a nonlinear fixed-charge transportation problem using a spanning tree-based genetic algorithm," *Computers & Industrial Engineering*, vol. 59, no. 2, pp. 259–271, 2010, doi: 10.1016/j.cie.2010.04.007.
- [27] S. Molla-Alizadeh-Zavardehi, M. Hajiaghayi-Keshteli, and R. Tavakkoli-Moghaddam, "Solving a capacitated fixed-charge transportation problem by artificial immune and genetic algorithms with a Prüfer number representation," *Expert Systems with Applications*, vol. 38, no. 8, pp. 10462–10474, 2011, doi: 10.1016/j.eswa.2011.02.093.
- [28] V. Balachandran and A. Perry, "Transportation type problems with quantity discounts," *Naval Research Logistics Quarterly*, vol. 23, no. 2, pp. 195–209, 1976, doi: 10.1002/nav.3800230203.

Appendix A. Instances and Reproducibility

This appendix documents every test instance used in the computational study: the three hand-specified instances I1–I3 in full numerical detail, the generation protocol for the six larger instances I4–I9, and the computational environment. Together with the released code, the information below suffices to regenerate every number reported in the paper.

A.1 Hand-Specified Instances I1–I3

Instances I1–I3 are specified directly rather than generated. Each table below reproduces the unit-cost matrix c_{ij} , with the supply vector a_i appended as the final column and the demand vector b_j appended as the final row. Fixed charges on I1–I3 follow the nominal regime $f_{ij} \sim U\{30, \dots, 120\}$ common to all nine instances and are regenerated by the released generator rather than reproduced here.

Table A.1 presents instance I1, a 3×4 problem. Total supply is 900 units against a total demand of 750, an excess supply of 150 units, so only the holding-cost vector $h = (4.5, 12.0, 7.8)$ is active and no shortage penalties are defined. The MILP benchmark certifies optimality here in 0.008 s, and the instance serves mainly as a correctness check for all pipelines.

Table A.1: Instance I1 (3×4): unit costs c_{ij} , supplies a_i , demands b_j ; $h = (4.5, 12.0, 7.8)$.

	$j = 1$	$j = 2$	$j = 3$	$j = 4$	a_i
$i = 1$	8	14	6	11	250
$i = 2$	10	17	9	5	300
$i = 3$	3	13	4	7	350
b_j	200	150	180	220	

Table A.2 presents instance I2 (5×5), which reverses the imbalance direction. Demand totals 1,108.8 units against a supply of 920, a shortfall of 188.8 units, so the shortage-penalty vector $p = (15.2, 32.1, 21.4, 45.6, 18.9)$ applies in place of holding costs. The fractional demands are exactly 1.12 times an integer base vector — a uniform 12% demand inflation. I2 also underlies the quantity-discount variant I2_QD.

Table A.2: Instance I2 (5×5): unit costs c_{ij} , supplies a_i , demands b_j ; $p = (15.2, 32.1, 21.4, 45.6, 18.9)$.

	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$	a_i
$i = 1$	6	7	11	5	14	180
$i = 2$	9	10	8	17	6	220
$i = 3$	4	13	15	9	11	150
$i = 4$	7	5	12	16	8	200
$i = 5$	18	3	10	14	19	170
b_j	224.0	201.6	179.2	235.2	268.8	

Table A.3 presents instance I3 (8×6), the largest hand-specified case. Supply totals 1,920.3 units against a demand of 1,220.3, an excess supply of 700 units — about 57% above demand — so the holding-cost vector $h = (3.2, 8.5, 12.1, 5.4, 9.8, 14.2, 6.7, 11.3)$ is active. The pronounced surplus forces every pipeline to trade holding costs off against routing on expensive arcs.

Table A.3: Instance I3 (8×6): unit costs c_{ij} , supplies a_i , demands b_j ;

$h = (3.2, 8.5, 12.1, 5.4, 9.8, 14.2, 6.7, 11.3)$.

	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$	$j = 6$	a_i
$i = 1$	5	12	8	19	3	15	215.3
$i = 2$	11	6	22	7	14	9	182.7
$i = 3$	8	18	4	16	20	11	301.5
$i = 4$	13	9	17	5	8	24	248.9
$i = 5$	6	21	10	14	12	7	165.4
$i = 6$	19	8	15	23	6	18	332.1
$i = 7$	4	16	9	11	25	13	198.6
$i = 8$	10	7	20	8	17	5	275.8
b_j	152.4	218.7	185.3	241.6	195.8	226.5	

A.2 Generated Instances I4–I9

Instances I4–I9 are generated by a single parameterized procedure driven by `numpy.random.RandomState(seed)`. Unit costs are integers drawn uniformly from $lo, \dots, hi$ per Table A.4. On the base side of the bipartition, the a_i or b_j values are integers from $U80, \dots, 120$; on the other side they are integers from $U70, \dots, 110$, normalized so that its total equals $1 + \text{ratio}$ times the base total. The normalized side is demand under excess demand and supply under excess supply, so the mode column of Table A.4 determines which penalty vector is active. Penalties are integers with $h_i \sim U\{2, \dots, 15\}$ and $p_j \sim U\{10, \dots, 50\}$.

Table A.4: Generation parameters for I4–I9 (ED = excess demand, ES = excess supply).

Instance	Size	Seed	Cost range	Mode	Ratio
I4	10×10	789	[1,30]	ED	0.20
I5	20×20	1011	[1,35]	ES	0.15

I6	30×30	1213	[1,40]	ED	0.15
I7	50×50	1415	[1,50]	ES	0.15
I8	100×100	1617	[1,50]	ED	0.20
I9	200×200	1819	[1,50]	ED	0.15

Two fixed-charge regimes are applied to the same underlying networks. The nominal regime draws $f_{ij} \sim U\{30, \dots, 120\}$; the realistic regime inflates the scale tenfold, $f_{ij} \sim U\{300, \dots, 1,200\}$, with the random stream re-initialized at seed+5,000 so that the draw is independent of the nominal one. Three quantity-discount variants — I2_QD, I7_QD, and I9_QD — keep the underlying instance data and replace the flat unit cost by the tiered rate $\gamma(x) c_{ij}$, with thresholds $\tau_2 = 0.25$ and $\tau_3 = 0.50$ of $M_{ij} = \min(a_i, b_j)$ and rates $r_2 = 0.85$, $r_3 = 0.70$. The discount is all-unit and non-convex, which is what destroys the LP structure for the exact solver.

A.3 Computational Environment and Code Availability

All linear and mixed-integer programs are solved with HiGHS through SciPy. The MILP benchmark uses `scipy.optimize.milp` (branch-and-bound with certified dual bounds) under a 120 s solver budget per instance, and the reported gap is $(\text{incumbent} - \text{dual})/\text{incumbent}$; the Proxy, Remedy, and Dummy pipelines solve their linear programs with the same SciPy/HiGHS stack. The ACO variants are implemented in pure NumPy with vectorized delta evaluation of the 4-cycle local search; the configuration is fixed at 30 ants, at most 100 iterations, $\alpha = 1$, $\beta = 2$, $\rho = 0.5$, and early stopping after 25 stagnant iterations. Unless stated otherwise, each stochastic configuration is run 10 times per instance with distinct seeds (3 runs on I8 under nominal fixed charges; 2 runs on I9_QD, reported descriptively), and the seed-inheritance ablation uses 5 runs per configuration. Dual-seeded ACO runs are deterministic (zero observed variance) and are run once. All times are wall-clock measurements on a single machine; where solver budget and wall time diverge — the 200×200 QD model takes about 19 s to build, so its 120 s budget corresponds to roughly 140 s wall — both figures are stated.

The complete instance generators (including all seeds), the raw result JSON files for every reported table, and the plotting scripts are available from the corresponding author upon reasonable request and will be deposited in a public repository upon acceptance of the paper.