Stable and Profitable Trading Platforms for Smallholder Commodity Supply Chains

Sergio Camelo
Stanford University, camelo@stanford.edu

Joann de Zegher
Massachusetts Institute of Technology, jfz@mit.edu

Dan A. Iancu
Stanford University, daniancu@stanford.edu

Abstract. Digital platforms that connect smallholder farmers with intermediaries offer a promising path to first-mile traceability and improved livelihoods. Yet such platforms face several challenges. Full disintermediation is often impossible because intermediaries are essential for logistics and deeply embedded in local informal relationship networks; stability is fragile, as farmers and intermediaries may revert to trading through local informal networks if they derive insufficient value on the platform; profit margins are tight; and operations are complex due to fragmentation, scale, heterogeneity, and significant uncertainty about local informal networks. To address these challenges, we develop a flexible model of the platform's joint decisions – matching farmers to intermediaries and setting payments – that captures potential off-platform deviations via local informal networks and enforces stability constraints that preclude them. Deviations are modeled with ambiguity sets that capture the platform's imperfect information and the breadth of intermediaries' relationship networks. We prove structural results and, leveraging them, design exact and approximate branch-and-bound algorithms. We pair a case study based on real data from Indonesia's palm-oil supply chain with a stylized version of the model that we solve analytically, to derive insights regarding the platform's main decisions. Our findings yield several managerial implications. Profitability depends critically on reducing transportation costs and obtaining at least partial data on informal local networks. Payments should be directed primarily to farmers; surprisingly, if intermediaries have *larger* relationship networks, platforms should increase payments to *farmers* and decrease payments to intermediaries, which is the most cost-effective way to ensure stability. In dealing with heterogeneous intermediaries, platforms should use efficient (i.e., minimum-cost) matchings unless some intermediaries have extremely large informal relationship networks, in which case they must be prioritized to ensure stability, despite a loss in efficiency.

Key words: smallholder farmers, commodity supply chains, matching platforms, distributionally robust optimization

1. Introduction

Digital platforms that facilitate transactions in the first-mile of agricultural value chains have proliferated in recent years. The growth has been driven by technology diffusion – expanded mobile-phone coverage and mobile-money solutions that lower communication and transaction costs (Aker 2010) – but also by rising demand for sustainable sourcing, as consumers and downstream buyers place greater emphasis on transparency and deforestation-free supply chains (Dong 2021, Warnes et al. 2025). Regulatory frameworks, most notably the EUDR (EU 2023) and the CSDDD (EU 2024), reinforce these pressures by requiring firms selling in the EU to ensure that commodities are produced on land not deforested since 2020 and that smallholder suppliers are paid in line

with living-income standards. Meeting these requirements calls for unprecedented traceability in commodity supply chains (Sodhi and Tang 2019, Warnes et al. 2025), with detailed data on input origins and first-mile transactions and tighter control over the farm-gate prices paid to farmers.

Digital platforms that match farmers with intermediaries and end-buyers, verify collections and drop-offs, and manage payment flows have thus emerged as a scalable infrastructure to deliver first-mile traceability and improve smallholder livelihoods. Illustratively, our field partner PemPem connects independent palm-oil smallholder farmers in Indonesia with intermediaries who have truck capacity to transport the fruit to local mills, and manages payments to both sides. Similarly, in Kenya, Zaidi Technologies (VC4A 2025) connects milk traders with farmers, purchases and processes the milk, and sells it in small shops in urban locations. And in Uganda, Kudu acts as an electronic clearing house designed to benefit smallholder farmers by matching them with buyers of agricultural produce (Bergquist et al. 2024).

This paper studies several interrelated challenges that threaten such platforms' ability to scale, using PemPem as a running example to illustrate many of the key issues.

The first is the *inability* to fully disintermediate. In some settings, platforms can connect smallholders directly to large buyers, using disintermediation to raise farm-gate prices and improve scale economies. But in contexts like the ones we consider, disintermediation is neither feasible nor desirable. In Indonesia, intermediaries–often former farmers–are embedded in local communities, where networks of relationships and trust are strong. They also provide an essential logistical service: they collect and transport fruit in areas with poor infrastructure (limited paved roads, dirt roads that deteriorate seasonally due to rain) and with extreme fragmentation across thousands of smallholders producing small quantities. These services are not readily substitutable and cannot be delivered at lower cost by other parties in the value chain. Moreover, platforms often find it easiest to onboard intermediaries first because they are the ones trading with processing mills, and to later onboard farmers using intermediary data or field work. In such cases, rather than outright disintermediation, platforms must connect farmers with existing intermediaries and work with – while also partly disrupting – informal trading networks.

A second challenge is *stability*: farmers and intermediaries may lack trust in newly launched platforms or the platform's value proposition may be insufficient, leading participants to revert to trading through preexisting local informal relationship networks. The challenge intensifies if business or legal constraints preclude the platform from paying intermediaries who are not assigned work, who could then "poach" farmers. Although platforms could promote stability through price

premiums or ancillary services that increase the value proposition, in many cases such incentives are not financially sustainable, as documented by our field partner and by existing literature (see §1.1), and the most effective lever remains a judicious matching of farmers and intermediaries to exploit economies of scale, combined with sufficient payments so that platform trade is preferred.

A third challenge is *profitability*. First-mile margins are often thin, and platforms may require significant cash inflows at inception that cannot be subsequently sustained, leading to failure. For instance, after entering multiple markets, Frubana withdrew from Brazil due to challenges in maintaining profitable operations (Aquino 2025). Similarly, eFishery, which provided technology and services to fish and shrimp farmers in Indonesia, exited the market citing limited sector margins as the primary driver (International Finance Correspondent 2025). And in India, DeHaat has grown rapidly but continues to operate at sustained losses (Kamal 2024). Such profitability pressures are compounded by the *stability* challenge–as large payments needed to attract or sustain trade quickly erode margins–and underscores the need for platforms to operate extremely efficiently.

A final challenge is *operational*. Efficient operations require accurate data and suitable algorithms to match farmers and intermediaries and design payments. Some data can be collected or estimated: scales distributed to farmers ensure quantity accuracy; GPS-enabled systems allow recording the exact location of farmer's plots and processing mills and estimating transit times and costs to ensure delivery without risking spoilage. But other data – most notably, reliable information on existing informal relationship networks – may be absent for newly established platforms or may even be strategically misreported. Such data gaps, together with the scale of the problem – with hundreds of farmers and dozens of intermediaries operating daily in a mill's supply shed – complicate operational decisions and also exacerbate both stability and profitability challenges.

This gives rise to several research questions. How should a newly launched first-mile commodity platform choose matches and payments to remain profitable while ensuring stable trade? Because paying both sides generously can quickly erode margins, whom should the platform pay more to ensure stability? Which intermediaries should be prioritized for a match – those with extensive informal networks (who might otherwise disrupt the platform) or those that are more cost-efficient (and could improve operations)? If informal networks were larger, how would payments and matches, the platform's profit, and the intermediary and farmer welfare change?

Our first contribution is to formalize a model that captures the key decisions and trade-offs facing a newly established platform and provides a flexible framework to address these questions. The platform jointly chooses farmer-intermediary matches and payments. Matches allow assigning the

same intermediary to multiple farmers to exploit economies of scale and must satisfy traceability constraints (collect all fruit when possible) and any relevant logistical constraints such as truck capacities. Payments are constrained so that only matched intermediaries can be compensated. Both matched and unmatched intermediaries may destabilize the platform by "deviating" from the platform's recommendation and transacting with farmers informally. When deviating, each intermediary samples farmers according to a probability distribution chosen from an *ambiguity set*. The platform only knows the ambiguity set–but not the distribution–and selects matchings and payments to maximize profit subject to stability, i.e., ensuring that recommended trades occur even under adversarial distributional choices. The ambiguity set's size therefore proxies both the platform's data quality and the intermediary's relationship network breadth/ability to deviate.

Our second contribution is to develop structural results and an algorithmic framework for solving the platform problem. We show that if the platform is overly conservative and assumes that any farmer-intermediary deviation is possible, it cannot earn positive profit while ensuring stability; accurately modeling a restricted set of deviations that reflects *more likely* relationships is therefore essential. We further show that, although matching and payment decisions are coupled, the coupling is weak: conditional on the set of intermediaries selected for matching, it is optimal to use the most cost-efficient matching and then set payments accordingly. Motivated by these insights, we propose a flexible scheme based on a set partitioning formulation and a branch-and-bound algorithm for solving the platform's problem, which supports many plausible data-driven specifications of ambiguity sets and logistical constraints. We develop both exact and approximate methods that exploit the problem structure to deliver significant speedups and that scale to realistic instances.

To address the remaining research questions, we combine a real-world case study with an analytical, stylized version of the model. The case study uses survey and GPS data from the Indonesian palm oil supply chain–typical of what a platform may have available in its early stages–to reconstruct informal intermediary-farmer relationship networks, estimate preferences and travel costs on paved and dirt roads, and compute realistic routing costs by solving vehicle routing problems. Guided by these data, we then consider a stylized version of the model that isolates the key tradeoffs and explains the mechanisms. The model assumes linear cost functions and normalizes farmer production to one unit of fruit, while retaining fixed costs to capture economies of scale, and heterogeneity in farmer visitation costs, intermediary cost-efficiency, and the size of each intermediary's informal network. As in the case study, the market is unbalanced: there are sufficient intermediaries to collect all fruit. Intermediaries have two types: a more numerous low type that

transacts with fewer farmers and is unable to fill a truck when deviating, and a smaller number of high-type intermediaries–with more extensive networks–who may fill a truck fully when deviating. Critically, either type may be more cost-efficient.

Drawing on the case study and the analytical model, we find that platform profitability hinges on two factors: its ability to increase efficiency by reducing transportation costs, and its ability to collect sufficient information to reduce ambiguity about relationship networks. Profits naturally decline as informal relationship networks expand because higher payments are needed to ensure stability, yet platforms can remain profitable and stable even under substantial ambiguity. When all relationship data are missing, sustaining operations would require an outside cash inflow of roughly 1% of the traded fruit value (in our case study).

We also find that most of the fruit value should accrue to *farmers* rather than to intermediaries or the platform itself. Surprisingly, if intermediaries' informal relationship networks were larger, this asymmetry would increase, so the platform should direct even *more* payments to *farmers*. The reason is that, when it is inefficient to match all intermediaries and unmatched intermediaries cannot be paid, allocating larger payments to farmers is the most cost-effective way to ensure stability.

Finally, although the platform generally matches intermediaries that are more cost-efficient or that have larger informal networks, subtler patterns emerge. If high-type intermediaries do not have extremely large networks, the platform prioritizes the most cost-efficient intermediaries and the resulting matches are efficient. But when high-type networks are very large, the platform must prioritize high-type intermediaries for matching to maintain stability, even at the expense of efficiency. Heterogeneity also shapes intermediary welfare: low-type intermediaries with smaller networks derive negligible welfare from the platform (even if they are efficient), whereas high-type intermediaries with large networks make positive welfare, which increases with their network size.

The results carry several practical messages for newly established platforms, summarized in §6.

1.1. Literature Review

Our paper relates to streams of work in operations management, operations research, and economics.

We contribute to the operations literature on agricultural supply chains in emerging economies (Kalkanci et al. 2019, Dong 2021). Closest to our work are studies of online marketplaces and platforms that intermediate agricultural supply (Ferreira et al. 2017), enable farm-equipment sharing by connecting farmers with booking agents (Adebola et al. 2025), provide transparent price information (Zhou et al. 2021, Shi et al. 2023), or allow auction designs that improve access and

prices in commodity exchanges (Levi et al. 2020, 2024). In contrast, we model informal relationship networks and the stability of trade—both essential for a platform's traceability function in commodity sourcing—and study matching and pricing as the platform's main levers. Our focus on traceability and first-mile logistics also links to blockchain systems (Dong 2021): while blockchains aim for end-to-end traceability via tamper-proof ledgers, their success hinges on accurate, verifiable first-mile data, which the digital trading platforms we study are designed to generate.

Our work also relates to the operations literature on ride-sharing, particularly one-to-many matchings where a single vehicle serves multiple customers. These problems are typically modeled as vehicle routing problems (VRPs); see Archetti and Bertazzi (2021) for a review. Variants relevant to smallholder value chains include milk-collection routing with limited road access (Belenguer et al. 2016) and time-window requirements to prevent spoilage (Hoogeboom et al. 2021). Our framework is general and can accommodate such constraints, provided that tractable subroutines (oracles) that compute minimum-cost matchings under these constraints are available.

Beyond matching, our work also incorporates pricing, which has been used in the ride-sharing literature to balance supply and demand (Bimpikis et al. 2019), to allocate transportation costs among passengers while accounting for heterogeneous preferences (Bian and Liu 2019), or to incentivize decisions that improve system efficiency, such as shifting demand between same-day and next-day deliveries (Banerjee et al. 2025) or steering users to favorable time windows (Waßmuth et al. 2023). In our setting, pricing instead ensures stability, guaranteeing that farmers and intermediaries prefer participating in the platform over maintaining informal relationships.

Our stability constraint resembles approaches used to model coalition-proofness in cooperative game theory. Similar concepts have been applied to cost sharing in transportation, including for the VRP (Engevall et al. 2004) and for Inventory Routing Problems (IRPs) (Özener et al. 2013); Guajardo and Rönnqvist (2016) provides a comprehensive review of cooperative game theory in transportation pricing. Most existing approaches rely on the core or, when the core is empty, on alternatives such as the nucleolus or the ϵ-core. These approaches are typically overly conservative because they assume all coalitions are feasible, which in our setting would make the platform profit negative (see Proposition 2). An alternative is to restrict feasible coalitions to a known network of relationships (Bilbao 2012) or to subsets of existing family structures (Warnes et al. 2025), but this requires (deterministic) information that may be difficult to obtain. Our model with ambiguity sets provides flexibility, covering the full range from one known coalition to all possible coalitions, along with intermediate cases where certain coalitions are more likely.

Our construction of ambiguity sets follows a distributionally robust optimization framework (Mohajerin Esfahani and Kuhn 2018). Robust and distributionally robust optimization has been applied in vehicle routing problems to account for uncertainty in demand (Gounaris et al. 2013, Subramanyam et al. 2020, Ardestani-Jaafari and Delage 2018), passenger availability (Hu et al. 2021), and travel times (Hoogeboom et al. 2021). But to our knowledge, these methods have not been applied to model uncertainty in informal relationships, which is critical in our context.

Our work also connects to literature on platform disintermediation, which studies mechanisms to prevent parties from bypassing the platform in subsequent transactions (Gu 2024). Prior contributions highlight contractual design, trust-building, and pricing instruments such as bonuses or promotions (Sekar and Siddiq 2023, Hagiu and Wright 2024). These approaches predominantly consider one-to-one matchings; by contrast, we study one-to-many matchings, which introduces distinct considerations related to complementarities in transportation costs, which make matching efficiency an important lever for retention, and stability, which must be addressed in a coalitional setting, with groups of actors (rather than pairs) deviating.

Lastly, our work relates to empirical economics literature on how search frictions and intermediation shape smallholder trading outcomes. A large-scale RCT of the Kudu mobile marketplace in Uganda documented the potential benefits of reducing search costs and leveraging scale economies (Bergquist et al. 2024), but complementary experiments have also found limitations: pure information interventions (e.g., SMS price services) were found to generate modest effects on farmer prices (Fafchamps and Minten 2012, Aker 2010); in Sierra Leone's cocoa chain, subsidies for traders primarily expanded advance payments to farmers, but without raising farm-gate prices (Casaburi and Reed 2020); in Kenya, platform entry yielded persistent intermediary markups and low pass-through of benefits to farmers (Bergquist and Dinerstein 2020). This partially motivates us to consider routine matching and pricing decisions as the main (indirect) levers for improving farmer welfare, and we find that profit-maximizing platforms with core concerns related to traceability should pay substantial prices to farmers in order to ensure stability of trade.

1.2. Notation

D^I denotes the set of vectors with entries indexed by elements of a finite index set I and taking values from a domain D; examples we use include $\mathbb{R}^I$, $\mathbb{R}^I_+$, and $\{0,1\}^I$. For $x \in D^I$ and $i \in I$, x_i denotes the component of x corresponding to index i. $\mathbf{1}_V \in \{0,1\}^I$ denotes the characteristic vector of the set $V \subseteq I$, i.e., a vector whose i-th component is 1 if $i \in V$ and 0 otherwise. We use $\mathbf{1}_i := \mathbf{1}_{\{i\}}$ and $\mathbf{1} := \mathbf{1}_D$ interchangeably. For $x, y \in D^I$ and $\alpha \in \mathbb{R}$, we use $x^\top y = \sum_{i \in I} x_i y_i$ and the shorthand

$x + \alpha$ for $x + \alpha\mathbf{1}$. We use "increasing", "decreasing," and other comparisons in a weak sense and say $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ is increasing if $f(y) \geq f(x)$ for any $x, y \in D$ with $y \geq x$. (f is "decreasing" if $-f$ is increasing.) $\Delta_I := \left\{\mathbb{P} \in \mathbb{R}^{\{0,1\}^I} : \mathbb{P} \geq 0, \ \sum_{s\in\{0,1\}^I} \mathbb{P}_s = 1\right\}$ denotes the set of all discrete probability distributions supported on $\{0,1\}^I$, $\text{supp}(\mathbb{P}) := \left\{s \in \{0,1\}^I : \mathbb{P}_s > 0\right\}$ denotes the support for $\mathbb{P} \in \Delta_I$, and $\delta_u \in \Delta_I$ denotes the distribution with unit mass on point $u \in \{0,1\}^I$. $\mathbf{1}(A)$ denotes the indicator/truth value of a logical condition A (equal to 1 if A is true and 0 otherwise).

2. Problem Formulation

A centralized platform matches farmers who produce fruit with intermediaries who collect and transport the fruit to a processing mill. The set $\mathcal{F}$ denotes all farmers, indexed by $f \in \mathcal{F}$, and the set $\mathcal{T}$ denotes all intermediaries with trucks, indexed by $t \in \mathcal{T}$. The platform's **matching** is given by a matrix $s \in \{0,1\}^{\mathcal{T}\times\mathcal{F}}$, where the t-th row $s_t \in \{0,1\}^{\mathcal{F}}$ is the **collection schedule** for intermediary $t \in \mathcal{T}$, with $s_{tf} = 1$ if and only if t collects the fruit of farmer f. If $s_t \neq 0$, we say that t is matched, and otherwise we say that t is unmatched. Each farmer f must be matched with exactly one intermediary, $\sum_{t\in\mathcal{T}} s_{tf} = 1$ for any $f \in \mathcal{F}$, but each intermediary t can be matched with several farmers. That each farmer is matched aligns with the platform's traceability priority of maximizing transactions; in the Indonesian palm oil supply chain, the supply of intermediaries with trucks is generally sufficient, making full collection feasible. Restricting each farmer to a single intermediary per harvest day reflects our historical data and standard practice, because farmers strongly prefer to sell to a single buyer per harvest due to the fixed costs of coordinating a collection and the fruit being perishable. (Our algorithms readily accommodate settings with unmatched farmers or multiple intermediaries assigned to partially collect from a farmer; see Appendix §D.1.)

Each farmer $f \in \mathcal{F}$ produces a quantity $q_f \geq 0$, assumed known. This is reasonable when farmers weigh their harvest before requesting collection and is consistent with several initiatives that provide farmers with scales to empower them and prevent conflicts with intermediaries. (Appendix §D.2 shows how to accommodate practical issues arising when quantities are uncertain.) Because farmers' costs are not impacted by the decisions we consider, we normalize them to zero for simplicity.

The collection schedule s_t for intermediary t is constrained to satisfy $s_t \in \mathcal{I}_t$, where $\mathcal{I}_t$ captures logistical constraints related to truck capacity or the intermediary's willingness to collect from specific farmers. When executing schedule s_t, intermediary t incurs a cost $c_t(s_t)$ that is monotonically increasing in s_t. The value $c_t(0)$ is not necessarily zero, which allows capturing a fixed cost (or opportunity cost) that t incurs when using his truck to collect fruit.

The centralized platform manages the payment flows by collecting the revenue from fruit sales at the processing mill and dividing it among farmers, intermediaries, and itself. The fruit is sold at the mill at per-unit price p. Each farmer $f \in \mathcal{F}$ receives a total payment $r_f \geq 0$ and each intermediary $t \in \mathcal{T}$ receives a total payment $z_t \geq 0$. Unmatched intermediaries receive zero payment: $z_t = 0$ for any $t \in \mathcal{T}$ with $s_t = 0$. This is an important business requirement that many commodity sourcing platforms – including our field partner – impose. (The constraint has important implications, which we discuss subsequently; but our algorithmic framework can readily accommodate settings where non-zero payments are possible.) The profit of a matched intermediary is $\pi_t = z_t - c_t(s_t)$ and the profit of an unmatched intermediary is $\pi_t = 0$. The platform's profit is:

$$\Pi = \sum_{f \in \mathcal{F}} p \cdot q_f - \sum_{f \in \mathcal{F}} r_f - \sum_{t \in \mathcal{T}} z_t.$$

Instead of accepting the platform's matches and payments, intermediaries can transact with farmers outside the platform ("deviate"). When deviating, intermediary t has access to some available farmers to choose from, which we model as a random vector u_t with realizations in $\{0,1\}^{\mathcal{F}}$, such that $u_{tf} = 1$ if and only if farmer f is willing to sell to intermediary t outside the platform. The **candidate vector** u_t follows a multivariate distribution $\mathbb{P}_t$ that is *chosen* by intermediary t from an ambiguity set of distributions $\mathcal{P}_t(\epsilon_t) \subseteq \Delta_{\mathcal{F}}$. This set is parameterized by an **ambiguity level** $\epsilon_t \geq 0$, with larger ϵ_t corresponding to a larger $\mathcal{P}_t(\epsilon_t)$. After observing the candidate vector u_t, the intermediary chooses any feasible subset of farmers from the available ones to deviate with; formally, t chooses a **deviation** $d_t \in \{0,1\}^{\mathcal{F}}$ satisfying $d_t \leq u_t$ and $d_t \in \mathcal{I}_t$. The intermediary pays each farmer f in a deviation the same amount r_f as the platform (which lowers the barrier to profitable deviations). In choosing the distribution $\mathbb{P}_t$ to sample candidates and the deviation d_t, intermediary t therefore obtains a maximum expected profit of $\hat{\pi}_t(r)$:

$$\hat{\pi}_t(r) := \sup_{\mathbb{P}_t \in \mathcal{P}_t} \left(\mathbb{E}_{u_t \sim \mathbb{P}_t} \left[\max_{d_t \in \mathcal{I}_t : d_t \leq u_t} \left((p \cdot q - r)^\top d_t - c_t(d_t) \right) \right] \right). \tag{1}$$

The platform ensures that no deviations occur by imposing the following **stability** constraints:

$$\pi_t = z_t - c_t(s_t) \geq \hat{\pi}_t(r) \qquad \forall t : s_t \neq 0 \tag{2a}$$

$$\pi_t = z_t - c_t(s_t) \geq 0 \qquad \forall t : s_t \neq 0 \tag{2b}$$

$$\pi_t = 0 \geq \hat{\pi}_t(r) \qquad \forall t : s_t = 0. \tag{2c}$$

Constraints (2a) and (2b) guarantee that any matched intermediary t prefers transacting through the platform rather than deviating or capturing his opportunity cost, respectively. (Note that (2b) and

the monotonicity of $c_t(\cdot)$ imply that $z_t \geq c_t(s_t) \geq c_t(0)$.) Constraint (2c) ensures that unmatched intermediaries cannot disrupt the platform through deviations.

The model of deviations based on ambiguity sets admits two practical interpretations. First, ambiguity sets can represent the strength of an intermediary's informal relationship network: with the probability $\mathbb{P}_t$ chosen by the intermediary, the set $\mathcal{P}_t$ and ambiguity level ϵ_t capture the breadth of his network or ability to attract business. Second, ambiguity sets can represent the platform's imperfect information about farmer-intermediary relations: first-mile platforms rarely observe the full network of bilateral preferences or transaction histories, so modeling distributions as unknown captures the quality of the data available. In practice, $\mathcal{P}_t$ and ϵ_t can be calibrated from available data to reflect the platform's risk tolerance (see §3 and §D.3 for several examples).

The platform's problem (PFP) is to find a matching s and payments r, z to maximize profit Π:

$$\begin{aligned}
\text{(PFP)} \quad & \underset{s\in\mathbb{R}^{\mathcal{T}\times\mathcal{F}}, r\in\mathbb{R}^{\mathcal{F}}, z\in\mathbb{R}^{\mathcal{T}}}{\text{maximize}} \ \Pi && && && \text{(3a)}\\
& \text{such that} \ \sum_{t\in\mathcal{T}} s_{tf} = 1, && \forall f \in \mathcal{F} && \text{(all farmers matched)} && \text{(3b)}\\
& && s_t \in \mathcal{I}_t, && \forall t \in \mathcal{T} && \text{(schedules are feasible)} && \text{(3c)}\\
& && z_t = 0, && \forall t \in \mathcal{T} : s_t = 0 && \text{(no pay if unmatched)} && \text{(3d)}\\
& && \text{(2a) – (2c)} && && \text{(stability).}
\end{aligned}$$

In practice, a commodity trading platform may face additional business constraints on its matching or pricing decisions – for example, requirements to use specific (transparent) pricing mechanisms, to prioritize certain farmers or intermediaries in the matching, to provide minimum income guarantees, or to incentivize repeated interactions. To focus the paper on the core tradeoff between profitability and stability, we do not model such features, but §B.5 briefly discusses them.

We seek optimal solutions for problem (PFP) and quantify their cost efficiency and the welfare they induce. We define the total cost, farmer welfare, and intermediary welfare, respectively, as

$$C(s) := \sum_{t\in\mathcal{T}: s_t \neq 0} c_t(s_t), \quad \mathcal{W}^{\mathcal{F}}(s, r, z) := \sum_{f\in\mathcal{F}} r_f, \quad \mathcal{W}^{\mathcal{T}}(s, r, z) := \sum_{t\in\mathcal{T}} \pi_t. \tag{4}$$

A solution (s, r, z) or a matching s is said to be **efficient** if it minimizes $C(s)$ among all matchings s that satisfy (3b)–(3c) (that is, efficient matchings are minimum-cost, feasible matchings). If problem (PFP) admits *multiple* optimal solutions, we examine whether any minimum-cost optima exist and consider the full range of values for $\mathcal{W}^{\mathcal{F}}$ and $\mathcal{W}^{\mathcal{T}}$ over optimal solutions.

3. Reformulations, Structural Results, and Tractable Algorithms

To derive structural results and propose tractable algorithms for (PFP), we impose additional assumptions on the feasible schedules $\mathcal{I}_t$ and costs $c_t(s_t)$ for each intermediary $t \in \mathcal{T}$.

ASSUMPTION 1. *For any $t \in \mathcal{T}$, the feasible schedules $\mathcal{I}_t$ form an independence system: if $s, s' \in \{0,1\}^{\mathcal{F}}$ with $s \leq s'$ and $s' \in \mathcal{I}_t$, then $s \in \mathcal{I}_t$. Moreover, there exists an oracle that, for any weights $w \in \mathbb{R}_+^{\mathcal{F}}$ and candidate vector $u \in \{0,1\}^{\mathcal{F}}$, solves in a reasonable time:*

$$\max_{d \in \mathcal{I}_t : d \leq u} \left[w^\top d - c_t(d) \right]. \tag{5}$$

The independence-system requirement means that if an intermediary can execute a schedule s', he can also execute any "easier" schedule $s \leq s'$. This allows capturing practical constraints such as truck capacity. The oracle condition ensures computational tractability and has a behavioral interpretation: intermediaries can "solve" their own deviation problem if offered a reward w_f per visited farmer, rather than relying on unrealistic mental models requiring intractable computations. Even if c_t does not satisfy this condition, one can substitute a *lower bound* cost function that does, which would yield a conservative reformulation of (PFP) whose solutions remain feasible under the true costs c_t. We next present examples of cost functions compatible with Assumption 1.

EXAMPLE 1 (LINEAR COSTS). Consider $c_t(s_t) = \sigma_t + \sum_{f \in \mathcal{F}} c_f s_{tf}$, where σ_t is a fixed cost and c_f the marginal cost of visiting farmer f. Interviews with intermediaries working with our field partner indicate a preference for this "mental model", which assigns each farmer a unique cost based on the amount of dirt road that must be traveled to reach the farmer. Solving (5) with truck capacity constraints entails a mixed-integer problem. We use this structure in our stylized model in §5.

EXAMPLE 2 (SUBMODULAR COSTS). A function c_t is submodular if $c_t(\mathbf{1}_{F \cup \{f\}}) - c_t(\mathbf{1}_F) \leq c_t(\mathbf{1}_{G \cup \{f\}}) - c_t(\mathbf{1}_G)$ for $G \subseteq F \subseteq \mathcal{F}$ and $f \notin F$. Submodularity captures scale economies, where the marginal cost of adding a farmer decreases with the set already visited. These functions provide a tractable family of models. If $\mathrm{conv}(\mathcal{I}_t)$ admits a polynomial-sized polyhedral representation—e.g., when $\mathcal{I}_t$ is a matroid (Schrijver et al. 2003)—then oracle (5) is a convex problem solvable with standard tools (Lovász 1983). With additional constraints, however, $\mathrm{conv}(\mathcal{I}_t)$ may lack such a representation; e.g., if $\mathcal{I}_t$ includes a knapsack constraint $q^\top s_t \leq K_t$ with unequal q, $\mathrm{conv}(\mathcal{I}_t)$ is the knapsack polytope. While not polynomially representable, outer and inner approximations can yield efficient bounds for which (5) remains tractable (see Hojny et al. 2020).

Example 3 (Vehicle Routing Costs). Transportation costs can be modeled as tours on a weighted complete graph $G = (N, E)$, where nodes N are farmer locations, depots, and the mill, and edge weights are travel costs. Then $c_t(s_t)$ is the minimum-cost tour from the depot, visiting $\{f : s_{tf} = 1\}$, delivering to the processing mill, and returning. The feasible set $\mathcal{I}_t$ may capture truck capacity or time-window constraints. Oracle (5) corresponds then to solving a Prize-Collecting Traveling Salesman Problem (PCTSP) (Balas 1989), which is NP-Hard on general graphs (Dror 1994), but tractable for special structures such as tree graphs, which we use in our case study in §4. High-quality heuristics and upper bounds also exist for general graphs.

Our next assumption concerns the ambiguity sets $\mathcal{P}_t$.

Assumption 2. *For every $t \in \mathcal{T}$, the ambiguity set $\mathcal{P}_t$ allows reformulating problem* (1) *as a finite-dimensional convex optimization problem whose separation oracle is exactly problem* (5).

As with Assumption 1, Assumption 2 ensures that each intermediary can "solve" his own deviation problem. The assumption is satisfied by many ambiguity sets calibrated from data; one example, which we leverage in §4 and §5, is the following.

Example 4 (Data-Driven Wasserstein Ambiguity). Assume the platform observes N past schedules $h_t^{(i)}, i = 1, \ldots, N$ for intermediary t. (For instance, such data can be obtained from GPS trackers on intermediaries' trucks.) We let $\hat{\mathbb{P}}_t$ be the empirical distribution of these schedules. A natural choice for $\mathcal{P}_t$ is a Wasserstein ambiguity set:

$$\mathcal{P}_t^{\mathbb{W}} = \left\{\mathbb{P}_t \in \Delta_{\mathcal{F}} : \mathbb{W}_w(\mathbb{P}_t, \hat{\mathbb{P}}_t) \leq \epsilon_t\right\}, \tag{6}$$

where $\epsilon_t \geq 0$ is a parameter that controls the size of the ambiguity set and $\mathbb{W}_w(\mathbb{P}_1, \mathbb{P}_2)$ is a weighted Wasserstein metric between two distributions $\mathbb{P}_1, \mathbb{P}_2 \in \Delta_{\mathcal{F}}$:

$$\mathbb{W}_w(\mathbb{P}_1, \mathbb{P}_2) := \min_{\gamma \in \Gamma(\mathbb{P}_1, \mathbb{P}_2)} \int_{\{0,1\}^{\mathcal{F}}} ||\xi_1 - \xi_2||_{1,w} d\gamma(\xi_1, \xi_2). \tag{7}$$

Here, $\Gamma(\mathbb{P}_1, \mathbb{P}_2)$ denotes the set of all couplings, i.e., joint distributions supported on $\{0,1\}^{\mathcal{F}} \times \{0,1\}^{\mathcal{F}}$ of the random variables ξ_1 and ξ_2 with marginal distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, respectively, and $||\cdot||_{1,w}$ is a weighted ℓ_1 norm: $||x||_{1,w} = \sum_{f \in \mathcal{F}} w_i |x_i|$ for some $w \in \mathbb{R}^{\mathcal{F}}$ with $w > 0$.

As discussed in §2, the ambiguity set $\mathcal{P}_t^{\mathbb{W}}$ and the ambiguity level ϵ_t bear two interpretations. One view is that ϵ_t measures an intermediary's relationship network and ability to deviate: if $w = q$ (as we use in §4 and §5), ϵ_t can be interpreted as the maximum quantity of fruit that the intermediary could collect (in expectation) from farmers with whom no relationships were observed in the data. The

limit $\epsilon_t \to \infty$ recovers all possible distributions, i.e., $\mathcal{P}_t^{\mathbb{W}} = \Delta_{\mathcal{F}}$, which is equivalent to intermediary t being able to deviate with any subset of farmers. A second view, based on the statistical guarantees provided by the Wasserstein distance, is that ϵ_t measures the accuracy of the platform's data: if $w = \mathbf{1}$ and $\hat{\mathbb{P}}_t$ is an empirical distribution of N samples drawn from an unknown distribution $\mathbb{P}_t^\star$, then $\mathcal{P}_t^{\mathbb{W}}$ contains $\mathbb{P}_t^\star$ with confidence $1-\beta$ when $\epsilon_t = O(\log(1/\beta)/N)$ (Fournier and Guillin 2015), so ϵ_t provides a measure of confidence that the platform has in its data.

The following proposition shows that the Wasserstein ambiguity set $\mathcal{P}_t^{\mathbb{W}}$ satisfies Assumption 1.

Proposition 1 **(Wasserstein)**. *If $\mathcal{P}_t = \mathcal{P}_t^{\mathbb{W}}$ defined in* (6)*, the optimal profit $\hat{\pi}_t(r)$ in* (1) *can be obtained by solving the following convex optimization problem in variables η and $\{\kappa^u\}_{u\in\operatorname{supp}(\hat{\mathbb{P}}_t)}$:*

$$\hat{\pi}_t(r) = \inf_{\eta\geq 0,\kappa^u} \left(\eta\cdot\epsilon_t + \sum_{u\in\operatorname{supp}(\hat{\mathbb{P}}_t)} \hat{\mathbb{P}}_t(u)\cdot\kappa^u \right) \tag{8a}$$

$$\text{s.t.} \sum_{f\in\mathcal{F}} \left(p\cdot q_f - r_f - \eta\cdot w_f\cdot \mathbf{1}(u_f=0)\right) d_{tf} - c_t(d_t) \leq \kappa^u \quad \forall d_t\in\mathcal{I}_t,\ \forall u\in\operatorname{supp}(\hat{\mathbb{P}}_t). \tag{8b}$$

Moreover, constraint (8b) *can be separated with $|\operatorname{supp}(\hat{\mathbb{P}}_t)|$ calls to oracle* (5)*.*

Constructing Wasserstein ambiguity sets requires some partial knowledge of farmer-intermediary relationships. When such data are scarce or unreliable, the platform may adopt the most conservative assumption, allowing for all possible deviations ($\epsilon_t \to \infty$). Although our model remains valid, the following result shows that the platform can never be profitable in this case.

Proposition 2 **(All deviations)**. *If every intermediary can deviate with any subset of farmers, i.e., $\mathcal{P}_t = \{\delta_d : d\in\mathcal{I}_t\}$ for every $t\in\mathcal{T}$, then the platform is not profitable, $\Pi^\star \leq 0$.*

This emphasizes the inherent tradeoff that the platform faces between profitability and stability and the need to collect additional data and reflect that in the construction of the ambiguity sets.

Many other data-driven ambiguity sets satisfy Assumption 2. Appendix §D.3 provides two examples: sets based on ϕ-divergences and sets based on known marginal distributions (which can be easily calibrated from machine learning models that estimate the likelihood of bilateral transactions), but many other examples from the DRO literature are compatible with the framework.

3.1. Reformulations and Structural Results

Assumptions 1 and 2 allow us to rewrite (PFP) as a compact Mixed Integer Convex program using a set-partitioning formulation. For each intermediary t and feasible schedule $s\in\mathcal{I}_t$, define $x_{ts}\in\{0,1\}$ as a binary variable indicating whether t performs schedule $s\in\mathcal{I}_t$, where $x_{t0}=1$ indicates that t is left

unmatched. Moreover, instead of the payment z_t, consider as decision variable the intermediary's profit π_t. With this, we rewrite (PFP) as the following equivalent optimization problem:

$$\text{(PFP)}_2 \qquad \underset{x,\, r,\, \pi}{\text{maximize}} \left(\sum_{f \in \mathcal{F}} p \cdot q_f - \sum_{f \in \mathcal{F}} r_f - \sum_{t \in \mathcal{T}} \pi_t - \sum_{t \in \mathcal{T}} \sum_{s \in \mathcal{I}_t,\, s \neq 0} x_{ts} \cdot c_t(s) \right) \tag{9a}$$

$$\text{s.t.} \sum_{t \in \mathcal{T}} \sum_{s \in \mathcal{I}_t : s_f = 1} x_{ts} = 1, \quad \forall f \in \mathcal{F} \tag{9b}$$

$$\sum_{s \in \mathcal{I}_t} x_{ts} = 1, \quad \forall t \in \mathcal{T} \tag{9c}$$

$$\pi_t \geq \hat{\pi}_t(r) \quad \forall t \in \mathcal{T} \tag{9d}$$

$$(1 - x_{t0}) M \geq \pi_t \quad \forall t \in \mathcal{T} \tag{9e}$$

$$\pi, r \geq 0, \quad x_{ts} \in \{0, 1\} \, \forall t \in \mathcal{T}, \forall s \in \mathcal{I}_t. \tag{9f}$$

(PFP)_2 is equivalent to (PFP) in that there is a one-to-one mapping between their respective optimal solution sets. To see this, note that (9b)-(9c) exactly mirror the matching constraints (3b)-(3c) (here, an intermediary is unmatched if and only if $x_{t0} = 1$). (9e) uses the big-M method to reformulate constraint (3d) that unmatched intermediaries are not paid, which requires $\pi_t = 0$. (M can be set as $\sum_{f \in \mathcal{F}} p \cdot q_f$). Lastly, the stability constraints (2a)-(2c) are equivalent to constraints (9d)-(9f).

Examining (PFP)_2 reveals an important structural property: the only coupling between the matching decisions x_{ts} and the pricing decisions r, π arises through constraints (9e), requiring zero pay for unmatched intermediaries. Because these constraints only concern *which intermediaries* are matched but not *with whom* they are matched, a natural decomposition of (PFP)_2 arises: conditional on which intermediaries are matched, the optimal matching minimizes the resulting costs.

Proposition 3. *If $(r^\star, \pi^\star, x^\star)$ is an optimal solution for (PFP)_2 and $y_t^\star = 1 - x_{t0}^\star$ indicates if intermediary t is matched, then $x^\star$ is a minimum-cost matching in a problem where intermediaries must be matched according to $y^\star$, that is, $x^\star$ is optimal in the following problem for $y = y^\star$:*

$$C(y) := \min_x \sum_{t \in \mathcal{T}} \sum_{s \in \mathcal{I}_t, s \neq 0} x_{ts} \cdot c_t(s) \tag{10a}$$

$$\text{s.t. (9b), (9c)} \tag{10b}$$

$$x_{t0} = 1 - y_t \quad \forall t \in \mathcal{T} \tag{10c}$$

$$x_{ts} \in \{0, 1\} \quad \forall t \in \mathcal{T}, \forall s \in \mathcal{I}_t. \tag{10d}$$

The first important implication of Proposition 3 is that $(\text{PFP})_2$ can be rewritten and interpreted as a two-stage optimization problem. Specifically, in the first stage, one should solve the problem:

$$(\text{PFP})_3 \qquad \underset{y,r,\pi}{\text{minimize}} \left(\sum_{f\in\mathcal{F}} r_f + \sum_{t\in\mathcal{T}} \pi_t + C(y) \right) \tag{11a}$$

$$\text{s.t. } \pi_t \geq \hat{\pi}_t(r) \quad \forall t \in \mathcal{T} \tag{11b}$$

$$y_t \cdot M \geq \pi_t \quad \forall t \in \mathcal{T} \tag{11c}$$

$$r, \pi \geq 0, \quad y_t \in \{0,1\}\, \forall t \in \mathcal{T} \tag{11d}$$

to find the vector $y^\star \in \{0,1\}^{\mathcal{T}}$ of matched intermediaries and the payments $r^\star$ and $\pi^\star$, respectively; in the second stage, one would then recover an optimal matching of intermediaries to farmers $x^\star_{ts}$ by solving (10) for $y = y^\star$. Subsequently and when no confusion can arise, we refer to $y \in \{0,1\}^{\mathcal{T}}$ as a matching with the understanding that this refers to the optimal solution in (10) for the given y.

Second, Proposition 3 highlights the critical role of constraints (3d) in (PFP) that require zero pay for unmatched intermediaries (there are (9e) in $(\text{S-PFP})_2$ and (11c) in $(\text{S-PFP})_3$, respectively). Without these constraints, the matching and pricing decisions are decoupled and any efficient (i.e., feasible, minimum-cost) matching is optimal overall in these problems, as stated in the next result.

COROLLARY 1. *For problem* (PFP) *without constraints* (3d)*, any efficient matching* $s \in \{0,1\}^{\mathcal{T}\times\mathcal{F}}$ *is optimal. Equivalently, for problem* $(\text{PFP})_3$ *without constraints* (11c)*, the optimal matching* $x^\star_{ts}$ *can be obtained by solving problem* (10) *without constraints* (10c) *(and with* y *arbitrary).*

Lastly, Proposition 3 also yields a simple approach for upper-bounding the value of $(\text{PFP})_3$: relax constraints (11c) and solve the matching and pricing problems separately. We exploit this next.

3.2. An Exact Algorithm

We propose a branch-and-bound algorithm for solving $(\text{PFP})_3$ exactly. The algorithm uses a binary tree where each node corresponds to a partial assignment of intermediaries as matched or unmatched. We first make an assumption that ensures that subproblems at each node are tractable.

ASSUMPTION 3 **(Matching Oracle)**. *With* $C(y)$ *defined in* (10)*, there exists an oracle that solves the following problem to optimality in "reasonable time" for any disjoint subsets* $Y_0, Y_1 \subseteq \mathcal{T}$*:*

$$M(Y_0, Y_1) := \min_{y\in\{0,1\}^{\mathcal{T}}} \Big\{ C(y) \;:\; y_t = b \quad \forall t \in Y_b,\ b \in \{0,1\} \Big\}. \tag{12}$$

The oracle in (12) computes a minimum-cost matching that *constrains* intermediaries in Y_0 to be unmatched and intermediaries in Y_1 to be matched (with all other intermediaries $\mathcal{T} \setminus (Y_0 \cup Y_1)$ unconstrained.) Depending on the costs c_t and feasible constraints $\mathcal{I}_t$, the oracle specializes to well-studied problems: with linear costs and truck capacity constraints, it reduces to a Generalized Assignment Problem (Ross and Soland 1975); with submodular costs, it resembles facility allocation with submodular functions (Svitkina and Tardos 2010); and with vehicle routing costs, it becomes a VRP. Because these problems are generally NP-hard, in §3.3 we describe how to relax Assumption 3 to design heuristics for $(\text{PFP})_3$ that can be solved efficiently in practice.

To describe our exact algorithm, let $G = (N, E)$ denote a binary tree rooted at $n_\emptyset = (\emptyset, \emptyset)$. Each node $n = (Y_0^n, Y_1^n)$ consists of two disjoint subsets of $\mathcal{T}$: Y_0^n denotes intermediaries required to be unmatched and Y_1^n denotes intermediaries required to be matched, and we allow for $Y_0^n \cup Y_1^n \subset \mathcal{T}$. We define the subproblem $(\text{PFP})_3^n$ in node n as the problem $(\text{PFP})_3$ with the additional constraint:

$$y_t = b \quad \forall t \in Y_b^n,\ \forall b \in \{0, 1\}.$$

The oracle in Assumption 3 enables us to compute upper and lower bounds for $(\text{PFP})_3^n$. For the lower bound, we solve $(\text{PFP})_3^n$ with the no-payment constraints (11c) enforced for $t \in Y_0^n \cup Y_1^n$ and removed for $t \in \mathcal{T} \setminus (Y_0 \cup Y_1)$. This relaxation decouples matching and pricing; the resulting optimal matching corresponds to $M(Y_0^n, Y_1^n)$, which we denote by y^n, and the optimal value yields the lower bound LB^n. This follows from Corollary 1, except that we apply it after fixing the matching according to Y_0^n and Y_1^n. For the upper bound, we solve $(\text{PFP})_3^n$ with the matching fixed to $y = y^n$, yielding UB^n. By Assumptions 1 and 2, both problems are convex and can be solved via row generation and oracle (5).

We now describe the branching procedure, which is initialized at the root node $n_\emptyset = (\emptyset, \emptyset)$ with bounds $\text{UB} = \infty$ and $\text{LB} = 0$. At each iteration a node n is selected, from a queue prioritized by the optimality gap on each node, and its bounds LB^n, UB^n are computed. If $\text{LB}^n > \text{UB}$, the node is pruned. Now, if $\text{LB}^n = \text{UB}^n$, then $(\text{PFP})_3^n$ is solved to optimality. Otherwise, y^n must violate some constraint (11c) which was previously relaxed, that is, there must exist an intermediary $\tau^n \in \mathcal{T} \setminus (Y_0^n \cup Y_1^n)$ with $y_{\tau^n}^n = 0$ and $\pi_{\tau^n}^n > 0$. We choose such intermediary with the largest violation $\pi_{\tau^n}^n$ and we branch by creating two children:

$$n_0 = (Y_0^n \cup \{\tau^n\}, Y_1^n), \qquad n_1 = (Y_0^n, Y_1^n \cup \{\tau^n\}), \tag{13}$$

which are added to the queue. If $\text{UB}^n < \text{UB}$, we update UB according. As is standard in branch-and-bound algorithms, we define a node as *active* if explored but with children still in the queue,

and after exploring n, the global lower bound LB is updated to be the minimum LB^n over all active nodes. The procedure finishes when no more nodes in the queue are left.

An optimality proof follows from standard branch-and-bound arguments, and is omitted.

We note that two design choices accelerate our algorithms. First, unlike the standard approach for weakly coupled MIPs, which would apply Lagrangian relaxation to the coupling constraints (11c) and require multiple oracle calls to (12) at each node, our implementation needs only a single oracle call per node of the branching tree. Moreover, if a node has descendants, at least one child reuses the same oracle call as its parent; in our case study, this further reduces runtime by at least a factor of five compared to the Lagrangian relaxation benchmark.

Second, we guide the branching in each node using the largest violation of the no-payment constraints (12); in the event of degeneracy, we break ties by selecting the solution that minimizes intermediary welfare $\mathcal{W}^{\mathcal{T}}$, which reduces penalties and ensures that branching occurs based on maximum violation. This rule delivers nearly a twofold speedup in our case study relative to the alternative with random tie-breaking and that ignores degeneracy.

3.3. Faster Heuristics

At each node of the branch-and-bound tree, computing bounds requires solving oracle (12), which can be costly depending on c_t. We outline three heuristics that provide valid upper bounds for $(\text{PFP})_3$ while reducing computational requirements.

Approximate matching oracle. Rather than solving oracle (12) exactly, we may use a γ-approximate solution $\tilde{y}$ with $C(\tilde{y}) \leq M(Y_0, Y_1) + \gamma$. Because the objective in $(\text{PFP})_3$ depends additively on $C(y)$, the resulting solution would be within γ of the optimum. This is especially useful when $C(y)$ involves hard combinatorial problems such as vehicle routing, where solving the oracle exactly is difficult but high-quality heuristics and approximation algorithms are available.

Minimum-cost matchings. If the gap between the upper bound $\text{UB}^{n_\emptyset}$ and lower bound $\text{LB}^{n_\emptyset}$ at the root note $n_\emptyset$ is small, the procedure can be stopped at the root, with a feasible solution derived from the upper bound. This entails a single call to oracle (12) and a solution based on an efficient matching (with pricing adjusted accordingly), which may offer a sufficiently good option in practice.

Lower bounds on costs. Solving each node requires multiple calls to oracle (5). If a convex lower bound for c_t is available, constraint (11b) can be dualized and solved directly. For instance, with a piecewise-linear lower bound on c_t, problem $(\text{PFP})_3$ reduces to a small linear program.

4. Case Study

We evaluate our framework through a case study of Indonesia's palm oil supply chain. Using proprietary data from the Riau region, we reconstruct the operations of intermediaries and farmers and evaluate outcomes under varying conditions.

Observed instances: Our data covers a three month period starting in June 2020. Combining GPS data from trucks operated by palm-oil intermediaries with machine-learning algorithms, we identify the farmers served by each intermediary and the quantities collected (see §B.1 for details). From these data, we construct 14 daily instances for the time window August 27 - September 9, 2020. On each day d, we let $\mathcal{T}$ denote the intermediaries available to collect fruit, and $\mathcal{F}$ the farmers who harvested fruit. Across all 14 instances, we have $|\mathcal{T}| = 14$, and $|\mathcal{F}|$ ranges between 15 and 40. We refer to these as the *observed instances*.

Synthetic instances: To assess robustness, we also generate *synthetic* instances by introducing controlled randomness in three dimensions: (i) fixed opportunity costs for intermediaries σ_t (defined below), (ii) historical farmer-intermediary relationships, and (iii) harvested quantities q_f. All random variables are drawn from distributions fitted to the observed data, ensuring that synthetic instances remain realistic. Details for the data-generation process are provided in §B.3.

Transportation costs and feasible schedules: For intermediary t on day d, we allow for all schedules $s_t \in \{0,1\}^{\mathcal{F}}$ that respect a truck capacity of nine tons: $\mathcal{I}_t = \left\{s_t \in \{0,1\}^{\mathcal{F}} : \sum_{f\in\mathcal{F}} s_{tf} q_f \leq 9.0\right\}$. The cost in executing schedule $s \in \mathcal{I}_t$ is $c_t(s) = \sigma_t + R_t(s)$, where σ_t is a fixed cost (interpreted as the rental value of the truck, elicited from surveys) and $R_t(s)$ is the cost of an optimal routing (a minimum-cost tour that departs the processing mill, visits all farmers with $s_f = 1$, and returns to the mill). To simplify computations, we approximate the road network by a tree graph G that spans farmer and processing mill locations. This assumption is motivated by driving patterns: intermediaries typically rely on paved roads, which resemble a tree network, and only deviate briefly on dirt roads to access individual plantations. Figure 1 illustrates the tree approximation overlaid on the real road network in the region of interest. (More details on calibration are provided in §B.2.)

Access to oracles: Under the tree graph G, Oracle 5 reduces to a Prize-Collecting Traveling Salesman Problem with capacity constraints, which is solvable given the tree structure via dynamic programming in pseudo-polynomial time $O(|V|K^2)$, where K is the truck capacity and $|V|$ is the number of vertices in the tree (Labbé et al. 1991). Oracle (12) reduces to a mixed-integer program, which we model and solve using Gurobi.

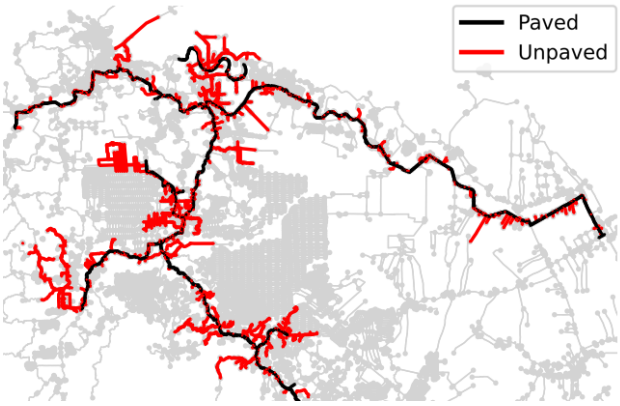

Figure 1 **Tree approximation of the underlying road network in the Riau region of Indonesia.**

Nominal distribution: For each intermediary t and each day, we construct the ambiguity set $\mathcal{P}_t$ around a nominal distribution $\hat{\mathbb{P}}_t = \delta_{\mathbf{1}_{\mathcal{H}_t}}$ that places all probability mass on the set $\mathcal{H}_t \subseteq \mathcal{F}$ of farmers historically observed to transact with t. Farmers are observed to transact with a single intermediary, so that $\{\mathcal{H}_t\}_{t\in\mathcal{T}}$ forms a partition of $\mathcal{F}$. In the data, most intermediaries maintain a narrow set of relationships and rarely fill their trucks, but a few have broader networks and operate near full capacity (see §B.4). This heterogeneity will play a central role in determining platform outcomes.

Ambiguity: We model uncertainty in relationships using the Wasserstein ambiguity set defined in Example 4, with weights $w_f = q_f$ that reflect the quantity of each farmer. This gives the ambiguity level ϵ_t an interpretation as the total fruit quantity that intermediary t is allowed to "add" beyond their historically observed relationships when deviating (Lemma 1 in §5 will formalize this).

Computational setup: We solve the (PFP) for both observed and synthetic instances on a computing cluster, running compute nodes with 16GB of RAM in parallel.

4.1. Ambiguity and Deviations

We examine how outcomes depend on the ambiguity levels ϵ_t. To start, we set $\epsilon_t = \epsilon$ for all $t \in \mathcal{T}$. Figure 2 plots platform profit (as a share of fruit value) against ϵ for both observed and synthetic instances. Profits decline steeply with greater ambiguity and are always negative at $\epsilon = 9$ tons, which is the smallest integer value ϵ at which the ambiguity set contains *all possible distributions*. At that value, the platform incurs an average loss of about 1%, which suggests that a 1% price premium on the traded fruit would suffice to sustain operations when intermediaries have powerful relationship networks or the platform's data are highly inaccurate.

We next examine how ambiguity affects the distribution of welfare between farmers and intermediaries. For each instance, we first calculate the optimal platform profit and then determine the

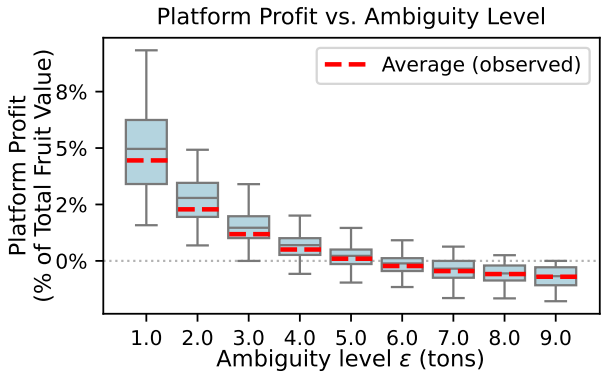

Figure 2 **Platform profit $\Pi^\star$ (as a percentage of total fruit value) as a function of the ambiguity level ϵ. Each boxplot summarizes 200 synthetic instances; the solid red line shows the average from the 14 observed instances.**

minimum and maximum welfare for farmers $\mathcal{W}^{\mathcal{F}}$ and intermediaries $\mathcal{W}^{\mathcal{T}}$ by optimizing over all platform-optimal solutions (because in many cases there is degeneracy). Figure 3 reports the minimum and maximum welfare for farmers and intermediaries in both observed and synthetic instances. Farmers consistently capture a larger share of total fruit value than intermediaries. Importantly, as ϵ increases, farmer welfare generally *increases* and intermediary welfare *decreases* (small non-monotonic fluctuations arise at larger values of ϵ, visible at 6 tons for the intermediary welfare.) These outcomes are surprising, particularly because ϵ is a proxy for each intermediary's ability to deviate, so the results suggest that a platform faced with more powerful intermediaries should increase payments *to farmers* in order to prevent deviations. Our stylized model in §5 will develop theory to explains these outcomes.

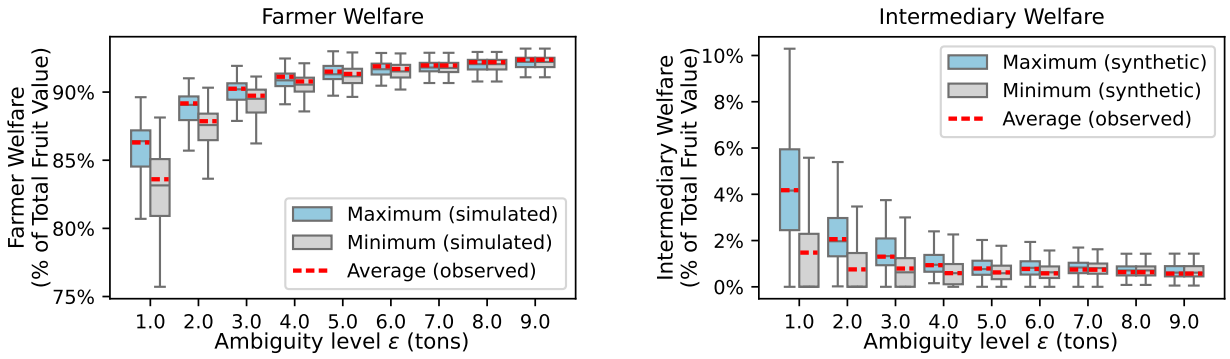

Figure 3 **Minimum and maximum welfare across optimal solutions for farmers (left) and intermediaries (right) as a function of ambiguity level ϵ. Boxplots summarize variation in synthetic instances; solid lines show averages from the 14 observed instances.**

4.2. The effect of heterogeneity

We study how two sources of heterogeneity affect outcomes: heterogeneity in the intermediaries' ability to deviate (the value of ϵ_t) and in their fixed costs σ_t. For these experiments, we focus on synthetic rather than observed instances, where we sample the ambiguity levels ϵ_t and the fixed costs σ_t independently for each intermediary (see §B.3 for details on sampling). To summarize an intermediary's "power to deviate," we define the *deviation potential* $\rho_t = \mathbb{E}_{u_t \sim \hat{\mathbb{P}}_t}\left[u_t^\top q\right] + \epsilon_t$, which corresponds to the maximum expected quantity that t can source in a deviation, from historically observed relationships and new, unobserved ones. (§5 will make this more precise.)

Figure 4 shows that intermediaries with low fixed costs σ_t and high deviation potential ρ_t are frequently matched and capture profits, whereas those with high fixed costs are excluded unless their deviation potential exceeds the truck capacity (9 tons), in which case they threaten the platform's stability and are matched and paid. Notably, intermediaries with deviation potential ρ_t below 5 tons make negligible profits, even when matched.

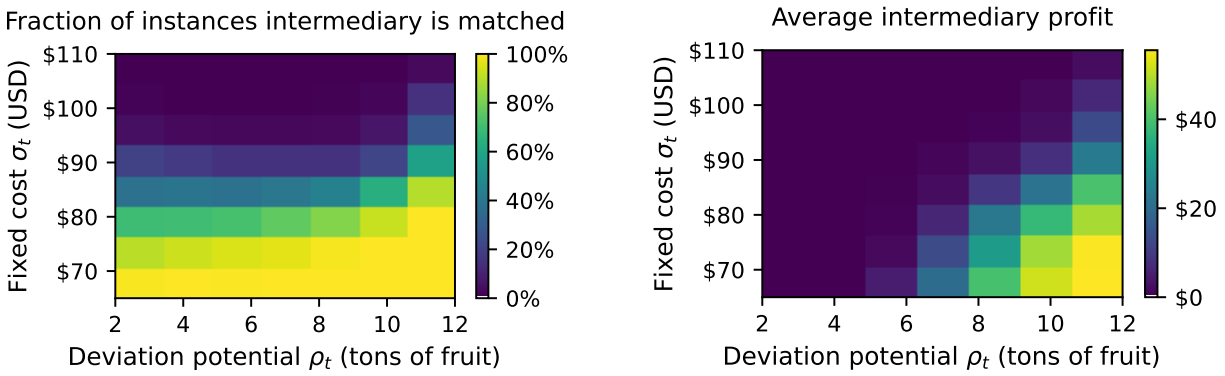

Figure 4 **Intermediary outcomes in synthetic instances with heterogeneous fixed costs σ_t and deviation potential ρ_t for optimal solutions that maximize intermediary welfare. Left: fraction of instances where intermediary is matched. Right: average profit of the intermediary.**

Finally, we analyze how payments to farmers are impacted by farmers' locations and the power of intermediaries they were observed working with historically. Figure 5 shows normalized (i.e., *per-ton*) payments in synthetic instances when farmers are grouped by the deviation potential of their known intermediary or by the required travel distance on dirt roads or paved roads needed to reach them (measured in a tour from the processing mill to the farmer and back). Results indicate that a profit-maximizing platform would strategically adjusts prices, paying more to farmers linked with more "powerful" intermediaries and/or lower road access costs. Because the former outcome

may raise concerns of equity in practice, Appendix §B.5 briefly examines the impact of using more transparent/fair pricing schemes on the platform's profit.

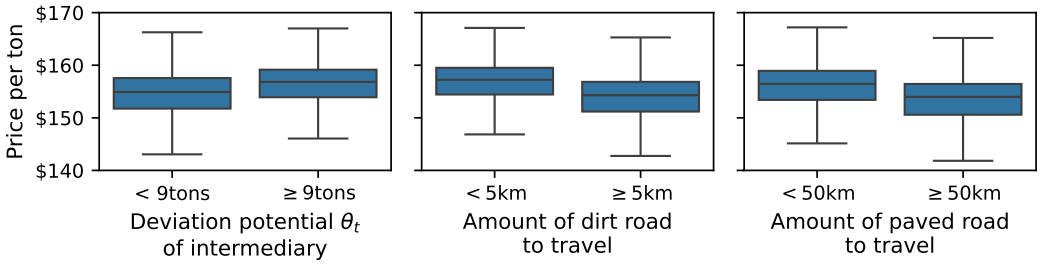

Figure 5 **Price paid to farmers (per ton) depending on the deviation potential ρ_t of the intermediary they transacted with historically (left) and the travel distance on dirt roads (center) and paved roads (right) to reach them.**

4.3. Scalable heuristics

Because intermediaries with lower fixed costs are preferred in matchings, we propose a simple heuristic for solving $(\text{PFP})_3$: fix the matching to a minimum-cost matching (i.e. the solution to $M(\emptyset, \emptyset)$ in (12)) and then optimize payments accordingly. Figure 6 evaluates performance in terms of computational runtime and optimality. With a four-hour time limit, this heuristic solves instances with up to 200 farmers, whereas the exact method only handles instances with at most 140 farmers. Because the number of farmers with daily harvests in one mill's supply shed rarely exceeds these values, the heuristic offers a practical, scalable solution for implementation. The potential caveat are the optimality gaps: although the median optimality gaps are generally below 2%, in some cases these could be as high as 8%, which could sufficiently hinder performance for a platform with razor-thin margins.

5. Stylized Model

We next formulate a stylized model to rationalize the empirical observations in the case study and generate insights on the platform's decisions. Each farmer $f \in \mathcal{F}$ produces one unit of fruit, $q_f = 1$. Each intermediary t owns a truck with integer capacity K. For simplicity, we take $|\mathcal{F}|$ as a multiple of K, so the minimum number of trucks necessary to collect from all farmers is $|\mathcal{F}|/K$. The cost function for intermediary $t \in \mathcal{T}$ when executing a collection schedule $s_t \in \{0,1\}^{\mathcal{F}}$ is $c_t(s_t) = \sigma_t + c^\top s_t$, where c_f is the marginal cost of collecting from farmer f and $\sigma_t > 0$ is a fixed cost. Feasible schedules must satisfy the truck capacity constraint: $\mathcal{I}_t = \left\{s_t \in \{0,1\}^{\mathcal{F}} : \mathbf{1}^\top s_t \le K\right\}$.

We model existing deviations using the Wasserstein ambiguity set defined in Example 4. The empirical distribution is $\hat{\mathbb{P}}_t = \delta_{\mathbf{1}_{\mathcal{H}_t}}$, consisting of one sample that corresponds to collecting from a

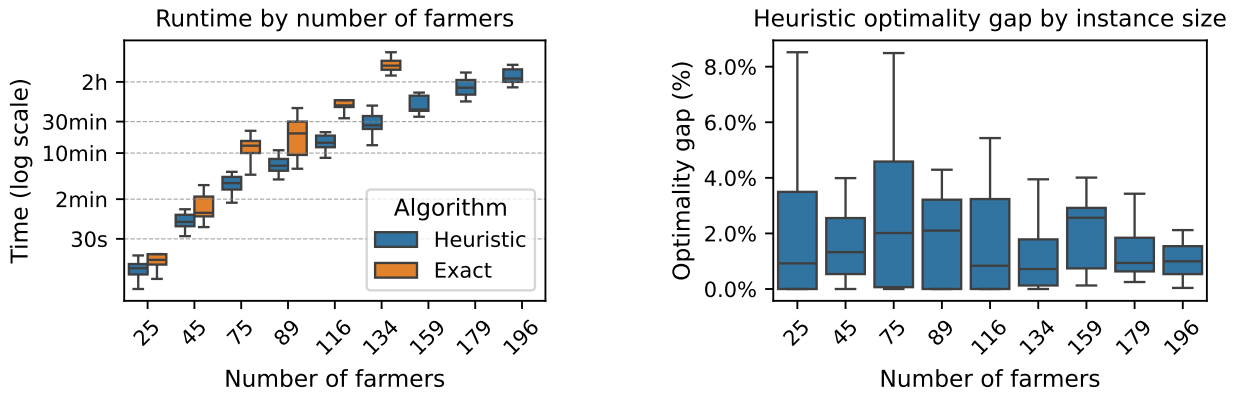

Figure 6 **Performance of the heuristic relative to the exact algorithm. Left: runtime. Right: optimality gap.**

set of farmers $\mathcal{H}_t \subseteq \mathcal{F}$ with whom intermediary t transacted historically. We refer to $\mathcal{H}_t$ as t's *status quo*. In line with our data, we assume that each farmer transacts with one intermediary, so $\{\mathcal{H}_t\}_{t\in\mathcal{T}}$ is a partition of $\mathcal{F}$. We define $\mathcal{H}_t^c := \mathcal{F} \setminus \mathcal{H}_t$ and $n_t := |\mathcal{H}_t|$. Because the status quo is feasible, we have $n_t \leq K$. The ambiguity set $\mathcal{P}_t^{\mathbb{W}}$ then consists of all distributions $\mathbb{P}_t$ such that $\mathbb{W}_1(\mathbb{P}_t, \hat{\mathbb{P}}_t) \leq \epsilon_t$, where $\mathbb{W}_1$ is the weighted Wasserstein distance defined in (7) with $w = \mathbf{1}$. We take the ambiguity level ϵ_t as a positive integer. Both n_t and ϵ_t are measures of intermediary t's relationship network, but in keeping with the terminology from §4, we refer to $n_t + \epsilon_t$ as the *deviation potential*, i.e., the total fruit quantity that t has access to when deviating.

For tractability, we limit the heterogeneity among intermediaries by considering two types given by $\theta \in \{h, \ell\}$. We partition the set $\mathcal{T}$ into $\mathcal{T}_h$ and $\mathcal{T}_\ell$, so that the fixed cost, status quo size, and ambiguity level of each intermediary t are type-specific: $\sigma_t = \sigma_\theta$, $n_t = n_\theta$, and $\epsilon_t = \epsilon_\theta$, for all $t \in \mathcal{T}_\theta$ and for each $\theta \in \{\ell, h\}$. This also induces a partition of farmers: $\mathcal{F} = \mathcal{F}_h \cup \mathcal{F}_\ell$, where $\mathcal{F}_\theta = \bigcup_{t\in\mathcal{T}_\theta} \mathcal{H}_t$ are all the farmers transacting with intermediaries of type $\theta \in \{\ell, h\}$ in the status quo.

We also make additional assumptions on the problem parameters, which we summarize next.

Assumption 4. *The parameters of the model satisfy the following conditions:*

(a) (Type ℓ has a smaller status-quo than h) $n_\ell < n_h$

(b) (Limited deviation potential) $\epsilon_\theta < n_h, \forall \theta \in \{\ell, h\}$

(c) (Type ℓ cannot fill truck when deviating) $n_\ell + \epsilon_\ell \leq K$

(d) (In deviations, type h is more cost-efficient than type ℓ) $\frac{\sigma_h}{\min(n_h+\epsilon_h, K)} < \frac{\sigma_\ell}{n_\ell+\epsilon_\ell}$

(e) (Type h collectively cannot cover market, but type ℓ can) $|\mathcal{T}_h|K < |\mathcal{F}|$ *and* $|\mathcal{T}_\ell|K > |\mathcal{F}|$

(f) (Commodity valuable) $\frac{\sigma_\ell}{n_\ell+\epsilon_\ell} + \max_{f\in\mathcal{F}} c_f < p$

(g) (No trivial cases) $\sigma_h \neq \sigma_\ell$, $n_\ell|\mathcal{T}_\ell| \geq K$ *and* $|\mathcal{T}_h| \geq 2$.

The requirements in Assumption 4 align with our palm-oil data (see §B.4) and streamline the analysis without sacrificing key insights. They are also intuitive. Assumptions (a)–(d) formalize how low-type intermediaries are "weaker" than high-types. Specifically, (a) states that low-types have fewer relationships in the status quo; (b) ensures deviations cannot generate more fruit from unknown farmers than high-type intermediaries collect from known farmers, limiting ambiguity. (c) assumes that low-types cannot fill their truck when deviating, which reflects reality and makes the problem interesting (otherwise, the setting becomes equivalent to $\epsilon_\ell = \epsilon_h = \infty$). (d) simplifies cases by making high-types more cost-effective when deviating. (e) assumes high-types cannot cover the whole market even at full capacity, whereas the more numerous low-types can; this reduces the number of cases without altering results qualitatively. (f) requires that the commodity be valuable enough for low-types to profitably include any farmer in a deviation, a mild strengthening of the natural assumption that intermediaries earn positive profits in the status quo. Finally, (g) ensures that inefficient matches exist ($\sigma_\ell \neq \sigma_h$) and avoids trivial cases. Importantly, the assumptions do not imply that high-types dominate uniformly: per (e), they cannot cover the market, and they may also face higher fixed costs ($\sigma_h > \sigma_\ell$). All results in §5 are derived under Assumption 4, which we do not restate for brevity.

We first reformulate (PFP) as the simpler problem of deciding the set of matched intermediaries T, the vector of farmer payments $r \in \mathbb{R}^{\mathcal{F}}$, and the vector of intermediary profits $\pi \in \mathbb{R}^{\mathcal{T}}$:

$$\text{(S-PFP)} \quad \max_{T \subseteq \mathcal{T}, r, \pi} \ \sum_{f \in \mathcal{F}} (p - r_f - c_f) - \sum_{t \in T} (\pi_t + \sigma_t) \tag{14a}$$

$$\text{s.t. } |T| \geq |\mathcal{F}|/K \tag{14b}$$

$$\pi_t \geq \hat{\pi}_t(r) = \sup_{\mathbb{P}_t \in \mathcal{P}_t^{\mathbb{W}}} \mathbb{E}_{u_t \sim \mathbb{P}_t} \left[\max_{d_t \leq u_t : d_t \in \mathcal{I}_t} \left(\left(p - r - c\right)^\top d_t - \sigma_t \right) \right] \ \forall t \in \mathcal{T} \tag{14c}$$

$$\pi_t = 0 \quad \forall t \in T^c \tag{14d}$$

$$\pi \geq 0, \quad r \geq 0. \tag{14e}$$

To see the equivalence, it is easiest to compare with $(\text{PFP})_2$, which is equivalent to (PFP). Any selection of intermediaries $T \subseteq \mathcal{T}$ in (S-PFP) satisfying (14b) yields a feasible matching by assigning each farmer to exactly one intermediary in T, with no more than K farmers assigned to the same intermediary. For such a match, the reformulation follows by considering the linear form of the cost function $c_t(s) = \sigma_t + c^\top s$ and that $\pi_t = 0$ for any unmatched intermediary $t \in T^c$. Subsequently, we refer to T as a "matching" with the understanding that this refers to any of the equivalent feasible

matchings induced by T, all of which have the same total transportation cost $\sum_{f\in\mathcal{F}} c_f + \sum_{t\in T}\sigma_t$. We also define the set of all *efficient* (i.e., feasible, minimum-cost) matchings as:

$$T^{\text{eff}} = \Big\{T \subseteq \mathcal{T} : |T| = |\mathcal{F}|/K,\ \mathcal{T}_h \subseteq T \text{ if } \sigma_h < \sigma_\ell \text{ and } T \subseteq \mathcal{T}_\ell \text{ if } \sigma_\ell < \sigma_h\Big\}. \tag{15}$$

We characterize optimal solutions for (S-PFP) and analyze their efficiency and the resulting farmer and intermediary welfare. Our first result offers a more intuitive reformulation: each deviating intermediary solves a generalized knapsack problem, and the platform decides which intermediaries to match and how much margin to extract from each farmer.

LEMMA 1. *For problem* (S-PFP) *defined in* (14)*,*

(i) the feasible set is unchanged if the ambiguity set $\mathcal{P}_t^{\mathbb{W}}$ is replaced with any of the following sets:

$$\mathcal{P}_t^1 = \left\{\mathbb{P}_t \in \Delta_{\mathcal{F}} \ :\ \mathbb{E}_{u_t\sim\mathbb{P}_t}\left[\sum_{f\in\mathcal{H}_t^c} u_{tf} + \sum_{f\in\mathcal{H}_t}(1-u_{tf})\right] \le \epsilon_t\right\}, \tag{16a}$$

$$\mathcal{P}_t^2 = \left\{\mathbb{P}_t \in \Delta_{\mathcal{F}} \ :\ \mathbb{E}_{u_t\sim\mathbb{P}_t}\left[\sum_{f\in\mathcal{H}_t^c} u_{tf}\right] \le \epsilon_t\right\}. \tag{16b}$$

(ii) intermediary t's optimal profit from deviating $\hat{\pi}_t(r)$ in (14c) *equals $\tilde{\pi}_t(v)$, where:*

$$\tilde{\pi}_t(v) := -\sigma_t + \max_{\mu_t\in[0,1]^{\mathcal{F}}}\left\{\mu_t^\top v \ :\ \sum_{f\in\mathcal{H}_t^c}\mu_{tf} \le \epsilon_t,\ \sum_{f\in\mathcal{F}}\mu_{tf} \le K\right\},\ \text{ and }\ v := p - r - c. \tag{17}$$

The optimal solution $\mu_t^\star$ to (17) *greedily sets $\mu_{tf}^\star = 1$ for farmers with $v_f > 0$, in decreasing order of v_f, so that at most ϵ_t farmers are selected from $\mathcal{H}_t^c$ and at most K are selected in total.*

(iii) with $\tilde{\pi}_t(v)$ defined as in (17)*, problem (*S-PFP*) has the same optimal value as problem:*

$$(\text{S-PFP})_2 \qquad \max_{T\subseteq\mathcal{T},\, v\in\mathbb{R}^{\mathcal{F}}} \sum_{f\in\mathcal{F}} v_f - \sum_{t\in T}\max\big(\tilde{\pi}_t(v), 0\big) - \sum_{t\in T}\sigma_t \tag{18a}$$

$$\text{s.t. } 0 \ge \tilde{\pi}_t(v), \quad \forall t \in T^c \tag{18b}$$

$$|T| \ge |\mathcal{F}|/K \tag{18c}$$

$$v_f \le p - c_f, \quad \forall f \in \mathcal{F}, \tag{18d}$$

and $(T^\star, v^\star)$ is optimal for (18) *if and only if $(T^\star, r^\star, \pi^\star)$ is optimal for* (S-PFP)*, where $r^\star = p - v^\star - c$ and $\pi_t^\star = \max\big(\tilde{\pi}_t(v^\star), 0\big)$ for any $t \in \mathcal{T}$.*

Part (i) provides an intuitive reformulation of deviations and reinterprets the ambiguity level ϵ_t. The set $\mathcal{P}_t^1$ shows that ϵ_t exactly acts as a budget on the *expected number of changes* that intermediary

t can make to his status quo $\mathcal{H}_t$ when sampling candidate vectors u_t. (Because $q_f = 1$, ϵ_t is also equivalent to the expected *quantity* changed relative to the status quo, in norm-1 sense.) Note that the expression under the expectation penalizes by one unit any new farmer added to the status quo, $u_{tf} = 1$ for $f \in \mathcal{H}_t^c$, and any farmer removed from the status quo, $u_{tf} = 0$ for $f \in \mathcal{H}_t$. The set $\mathcal{P}_t^2$ sharpens this by establishing that it is optimal to use the entire budget ϵ_t to add new candidate farmers to the status quo rather than remove from the status quo. The intuition is straightforward: candidate vectors u_t specify farmers *available* for deviations, and given u_t, intermediary t then selects d_t with $d_t \leq u_t$. A larger u_t benefits the intermediary, so it is optimal to use the entire budget ϵ_t for including farmers from $\mathcal{H}_t^c$ rather than removing farmers from $\mathcal{H}_t$.

Part (ii) reformulates the intermediary's deviation profit in (17). The decision variable $\mu_{tf} \in [0, 1]$ denotes the fraction of farmer f's unit production collected by intermediary t. When deviating, intermediary t earns a margin $v_f = p - r_f - c_f$ from farmer f, reducing his problem to a continuous knapsack with two constraints: at most ϵ_t units may be collected from farmers outside the status quo ($\mathcal{H}_t^c$), and total collection cannot exceed truck capacity K. The optimal policy is greedy: fully collect from farmers with strictly positive margins ($v_f > 0$), subject to the ϵ_t and K constraints.

Finally, Part (iii) reformulates the platform's problem as selecting a set of matched intermediaries $T \subseteq \mathcal{T}$ and margins $v \in \mathbb{R}^{\mathcal{F}}$. The formulation is equivalent in that a one-to-one mapping between the optimal solution sets exists. This is important when considering secondary objectives – such as farmer or intermediary welfare or efficiency – defined over all optimal solutions.

In view of Lemma 1-(iii), we focus on characterizing optimal solutions to $(\text{S-PFP})_2$. Our first result describes the platform's optimal profit $\Pi^\star$ and the optimal matching(s) $T^\star$.

PROPOSITION 4. *There exist constants A, B_ℓ, B_h and threshold $\underline{\epsilon}_h$ such that the optimal profit $\Pi^\star$ and the optimal matchings in* $(\text{S-PFP})_2$ *are as follows:*

(i) If $\sigma_h < \sigma_\ell$, then $\Pi^\star = A$, all high-type intermediaries are matched ($\mathcal{T}_h \subseteq T^\star$), and the set of optimal matchings is T^{eff}.

(ii) If $\sigma_h > \sigma_\ell$ and $n_h + \epsilon_h \leq K$, then $\Pi^\star = A$, only low-types intermediaries are matched ($T^\star \subseteq \mathcal{T}_\ell$), and the set of optimal matchings is T^{eff}.

(iii) If $\sigma_h > \sigma_\ell$, $n_h + \epsilon_h > K$, and $\epsilon_h < \underline{\epsilon}_h$, then $\Pi^\star = A - B_\ell$, only low-types intermediaries are matched ($T^\star \subseteq \mathcal{T}_\ell$), and the set of optimal matchings is T^{eff}.

(iv) If $\sigma_h > \sigma_\ell$, $n_h + \epsilon_h > K$, and $\epsilon_h > \underline{\epsilon}_h$, then $\Pi^\star = A - B_h$ and any optimal matching $T^\star$ includes all high-type intermediaries ($\mathcal{T}_h \subseteq T^\star$), satisfies $|T^\star| = |\mathcal{F}|/K$, and is not *efficient, $T^\star \notin T^{\text{eff}}$.*

Moreover, the optimal profit $\Pi^\star$ *is decreasing in* ϵ_ℓ *and* ϵ_h*, the values* $A, B_\ell, B_h, \underline{\epsilon}_h$ *are given by:*

$$A := \Delta C - \frac{\epsilon_\ell|\mathcal{T}_\ell| + \min(\epsilon_h, K - n_h)|\mathcal{T}_h|}{n_\ell + \epsilon_\ell}\sigma_\ell, \text{ where } \Delta C := \sum_{t\in\mathcal{T}} \sigma_t - \min_{\substack{T\subseteq\mathcal{T} \\ |T|\geq|\mathcal{F}|/K}} \sum_{t\in T} \sigma_t,$$

$$B_\ell := \min\left(|\mathcal{F}| - |\mathcal{T}_h|K,\ |\mathcal{T}_h|K \cdot \frac{n_h + \epsilon_h - K}{K - \epsilon_h}\right)\left(\frac{\sigma_\ell}{n_\ell + \epsilon_\ell} - \frac{\sigma_h}{K}\right), \qquad B_h := |\mathcal{T}_h|(\sigma_h - \sigma_\ell),$$

and $\underline{\epsilon}_h$ *is the solution to the equation* $B_\ell = B_h$*, which is increasing in* σ_h *and* ϵ_ℓ*, decreasing in* σ_ℓ*, and satisfies* $\underline{\epsilon}_h > K - n_h$*.*

The result illustrates how the heterogeneity in fixed costs σ_t and in ambiguity levels ϵ_t determines optimal outcomes. When $\sigma_h < \sigma_\ell$, high-type intermediaries dominate low-types in all regards, so the platform's optimal matchings utilize high-types to the fullest extend possible: any optimal matching $T^\star$ satisfies $\mathcal{T}_h \subseteq T^\star$ and all optimal matchings are efficient. When high-types intermediaries have higher fixed costs than low-types ($\sigma_h > \sigma_\ell$) but their ability to deviate is not very large ($\epsilon_h < \underline{\epsilon}_h$), they do not threaten the platform's stability, so the platform remains efficient and optimal matchings only rely on low-types (any optimal matching $T^\star$ satisfies $T^\star \subseteq \mathcal{T}_\ell$). Lastly, when high-type intermediaries have sufficient ability to deviate ($\epsilon_h > \underline{\epsilon}_h$), they can credibly threaten to destabilize the platform; to preserve stability, the platform must then prioritize matching the high-types–which allows paying them–so all optimal matches are *inefficient* and satisfy $\mathcal{T}_h \subseteq \mathcal{T}^\star$. These results echo the patterns in Figure 4, where intermediaries with high deviation potential $\rho_t = n_t + \epsilon_t$ are matched despite having high fixed costs, simply because their informal networks are too strong to ignore. Notably, as high-types become less efficient (larger σ_h) or as low-types become more efficient (smaller σ_ℓ) or more capable to deviate (larger ϵ_ℓ), the high-types must exhibit increasingly stronger ability to deviate to create threats leading to this regime (because the threshold $\underline{\epsilon}_h$ increases).

Proposition 4 also highlights how the platform's profit is bounded by its ability to reduce transportation costs. Specifically, note that ΔC is the maximum possible reduction in total transportation costs, and serves as an upper bound on the profit $\Pi^\star$. When ambiguity levels $\epsilon_h, \epsilon_\ell$ are small, deviations are limited, payments remain low, and $\Pi^\star$ approaches this bound. As ambiguity grows, the platform must raise payments to deter deviations, which reduces its optimal profit.

The next result characterizes the optimal welfare of farmers and intermediaries. We define $\mathcal{S}^\star$ as the set of optimal solutions $(T^\star, v^\star)$ for (S-PFP)$_2$ and $\underline{\mathcal{W}}^{\mathcal{A}}$ and $\overline{\mathcal{W}}^{\mathcal{A}}$ as the smallest and largest welfare, respectively, achieved in any optimal solution by $\mathcal{A} \in \{\mathcal{F}, \mathcal{T}\}$, that is,

$$\overline{\mathcal{W}}^{\mathcal{A}} = \max_{(T,v)\in\mathcal{S}^\star} \mathcal{W}^{\mathcal{A}}(v), \quad \underline{\mathcal{W}}^{\mathcal{A}} = \min_{(T,v)\in\mathcal{S}^\star} \mathcal{W}^{\mathcal{A}}(v),$$

where $\mathcal{W}^{\mathcal{F}}(v) = \sum_{f\in\mathcal{F}}(p - v_f - c_f)$ denotes farmer welfare and $\mathcal{W}^{\mathcal{T}}(v) = \sum_{t\in\mathcal{T}}\max(\tilde{\pi}(v), 0)$ denotes intermediary welfare, respectively, under a solution (T, v).

PROPOSITION 5. *The optimal welfare outcomes in* (S-PFP)$_2$ *are as follows:*

(i) Low-type intermediaries earn zero profit in all optimal solutions, $\pi_t^\star = 0, \forall t \in \mathcal{T}_\ell$. In cases (i) and (iv) of Proposition 4 (when high-types are matched), the largest intermediary welfare is

$$\overline{\mathcal{W}}^{\mathcal{T}} = |\mathcal{T}_h|\left(\min(K, n_h + \epsilon_h)\tfrac{\sigma_\ell}{n_\ell+\epsilon_\ell} - \sigma_h\right), \tag{19}$$

which increases in ϵ_h and σ_ℓ, decreases in ϵ_ℓ and σ_h, and decreases in ϵ if $\epsilon_h = \epsilon_\ell = \epsilon$. Otherwise (i.e., in cases (ii), (iii) of Proposition 4), intermediary welfare is zero, $\mathcal{W}^{\mathcal{T}} = 0$.

(ii) If cases (i) and (iv) of Proposition 4, the minimum farmer welfare $\underline{\mathcal{W}}^{\mathcal{F}}$ decreases in ϵ_h, increases in ϵ_ℓ, and increases in ϵ if $\epsilon_h = \epsilon_\ell = \epsilon$. In cases (ii) and (iii) of Proposition 4, $\underline{\mathcal{W}}^{\mathcal{F}}$ increases in ϵ_h and ϵ_ℓ. Moreover, the minimum farmer welfare is bounded below:

$$\underline{\mathcal{W}}^{\mathcal{F}} \;\geq\; \sum_{f\in\mathcal{F}}\left(p - c_f - \frac{\sigma_\ell}{n_\ell + \epsilon_\ell}\right). \tag{20}$$

(iii) There exists an optimal solution (T, v) with equal margins for all farmers working with intermediaries of the same type, i.e., $v_{f_1} = v_{f_2}$ for all $f_1, f_2 \in \mathcal{F}_\theta$ and $\theta \in \{\ell, h\}$. For such solutions, the welfare of a farmer working with a high-type intermediary is always higher than the welfare of a farmer working with a low-type intermediary,

$$r_{f_1} = p - c_{f_1} - v_{f_1} \;\geq\; r_{f_2} = p - c_{f_2} - v_{f_2} \quad \forall f_1 \in \mathcal{F}_h,\ \forall f_2 \in \mathcal{F}_\ell. \tag{21}$$

To interpret the results, recall that the platform's core strategy for maximizing profits hinges on reducing transportation costs, which requires leaving several intermediaries unmatched. Indeed, Proposition 4 shows that the number of matched intermediaries is always minimized, $|T^\star| = |\mathcal{F}|/K$. Because unmatched intermediaries cannot be paid, their deviation profit $\tilde{\pi}_t(v)$ must be held below zero to ensure stability. The platform achieves this by decreasing the margins v_f as much as possible, because $\tilde{\pi}_t(v)$ is monotonic in v. Then, as Proposition 5(i) describes, low-type intermediaries earn zero profit in all optimal solutions, even when matched, whereas high-type intermediaries make positive profits when matched, namely either when they are efficient or when they have very extensive deviation potential. The logic matches the patterns in Figure 4, where intermediary profits rise when fixed costs are low and/or deviation potentials ρ_t are high. The result also suggests that intermediary welfare rises with ϵ_h and σ_ℓ, but declines with ϵ_ℓ and σ_h; put differently, high types

benefit when heterogeneity is stark: when they are either much more cost-efficient or when they have much greater deviation potential than low-types.

Proposition 5 also highlights the interesting case that arises under symmetric ambiguity, $\epsilon = \epsilon_h = \epsilon_\ell$, when increasing ϵ–which strengthens the relationship networks of (all) intermediaries–*lowers* intermediary welfare. The reason is that a higher $\epsilon = \epsilon_\ell$ also increases the power to deviate for low-types, yet the platform cannot compensate low-type intermediaries because some must remain unmatched. To maintain stability, the platform optimally raises payments for *farmers*, and reduces the payments made to any matched high-types. This counterintuitive finding is well aligned with the empirical observation in Figure 3.

Considering the global effect of ϵ_h also reveals why intermediary welfare may have non-monotonic dependency on the (high-type) deviation potential. Specifically, in cases (iii) and (iv), high-types having lower deviation potential ($\epsilon_h < \underline{\epsilon}_h$) results in all intermediaries having zero welfare, whereas a sufficiently large high-type deviation potential ($\epsilon_h > \underline{\epsilon}_h$) leads to a strictly positive intermediary welfare because high-types are matched. In the latter regime, intermediary welfare becomes strictly positive, but then decreases in ϵ_h as this parameter rises. This explains the non-monotonic dependency in intermediary welfare documented in Figure 3, although the pattern is less prominent that the overall trend of intermediary welfare decreasing with ambiguity levels.

Farmer welfare exhibits the mirror image of these effects. As part (ii) states, greater ambiguity and deviation potential generally *increases* farmer gains, except when type-h intermediaries are matched and ϵ_h increases. In the symmetric case $\epsilon = \epsilon_h = \epsilon_\ell$, however, farmer welfare always increases with ϵ, again matching the empirical observations in Figure 3.

Equations (19) and (20) also reveal why the farmer welfare significantly dominates intermediary welfare. Note that intermediary welfare scales with the fixed costs σ_ℓ and deviation potentials $n_\theta + \epsilon_\theta$, whereas the (lower bound on) farmer welfare scales with the fruit price p. In realistic settings where p is much larger than σ_θ (such as in our case study), most of the surplus will then accrue to farmers. Intuitively, the reason why increases in p are effectively passed to farmers is that passing such increases to intermediaries would make their deviation profits large and require the platform to match them to ensure stability, which would undermine efficiency/cost savings.

Finally, (21) shows that there exist solutions in which a farmer's compensation depends solely on cost heterogeneity and the intermediary they work with in the status quo, and that farmers working with high-type intermediaries are compensated more than farmers working with low-type intermediaries. This is consistent with the empirical patterns in Figure 5, where farmers located

farther were typically paid less and payment differentials reflected the deviation potential of the intermediary servicing them in the status-quo.

6. Conclusions, Limitations, and Directions For Further Research

This work proposed an analytical framework to guide the decisions of a newly established first-mile commodity platform, while capturing limited available data on informal relationship networks and key logistical and business constraints. Using structural results, we showed that the problem is weakly coupled and we developed tractable exact and approximate algorithms for solving the platform's problem. Combining a case study using data from Indonesia with a stylized version of the model, we derived several practical insights.

Our results show that collecting data on informal relationship networks is critical. Without such data, a platform that ignores stability risks losing transactions to informal trade, while one that enforces stability under conservative assumptions must make large payments to farmers and intermediaries, requiring cash inflows or price premiums.

In markets with abundant intermediaries and trucking capacity, relationship data and cost-efficiency jointly determine which intermediaries the platform should work with. At inception, the platform should prioritize the most cost-efficient intermediaries, except when some have especially powerful relationship networks, in which case they must be prioritized to ensure stability.

The ability to pay unmatched intermediaries also shapes operations and efficiency. Platforms that can pay retainer or reservation fees to unmatched intermediaries would operate fully efficiently, hiring only the most cost-effective intermediaries and designing payments (to everyone) to ensure stability. In contrast, platforms unable to compensate unmatched intermediaries may need to allocate work inefficiently simply to maintain stability.

The results also matter for platforms concerned with smallholder welfare. Even if profit-maximizing, platforms should channel most of the value to farmers, far more than to intermediaries or themselves. Moreover, when powerful intermediaries threaten stability, directing larger payments to *farmers* is the most cost-effective way to stabilize operations and limit efficiency losses.

Our work also has limitations that point to important future research directions. First, our model focused on the decisions of a newly established/emerging platform and ignored how matching decisions may reshape the informal relationship networks over time. As platforms operate, intermediaries may discover new farmers outside their initial networks, expanding their scope for future deviations. The stylized model in §5 suggests that such dynamics could generate counterbalancing

effects: because some low-type intermediaries must always be matched, that would increase their networks over time; depending on whether (and with whom) high-types are matched, the heterogeneity across relationship networks may either increase or decrease, with important implications for stability and profitability. Future work could also consider other aspects of a repeated interaction, such as batching collections over consecutive days (when products are not perishable), or offering loyalty incentives for repeated transactions.

Second, to support onboarding, further research could explicitly model incentives for intermediaries to truthfully report their existing relationships and/or costs. Our results suggest that reporting incentives may partly be aligned, because more cost-effective intermediaries or those with stronger networks are prioritized for matching. To deter intermediaries from overstating their relationship networks, the platform can verify ties with farmers, as is common practice. Because farmers' welfare decreases with ϵ_h, farmers would not have incentives to support inflated reports of more powerful intermediaries' networks, but because their welfare increases with ϵ_ℓ, they may have incentives to misreport low-type networks. Modeling and studying such incentives carefully could shed more light on stability and guide the design of strategy-proof mechanisms, such as those using Moulin or acyclic mechanisms (Mehta et al. 2007).

Future work could also consider aspects of fairness or interpretability in the platform's decisions. For instance, the platform may impose minimum price or income guarantees, which are important for retention and stability (Gur et al. 2021), or may require simpler and transparent pricing schemes, e.g., based solely on observable farmer characteristics like location. Our framework accomodates such considerations, as we show §B.5, but future work could examine the issues in more depth.

Lastly, future work could also be devoted to testing different algorithmic approaches or heuristics that solve specific variations of the platform problem. For instance, *ng*-path/route relaxations (Baldacci et al. 2011), local branching (Fischetti and Lodi 2003), local search (Bertsimas et al. 2013), or other metaheuristics (Archetti and Speranza 2014) could be used to solve the problem to (near) optimality, while scaling to much larger instances.

References

Adebola O, Arora P, Zhang C (2025) Sharing platforms in emerging markets: The role of human intermediaries, sSRN 4190725.

Aker JC (2010) Information from markets near and far: Mobile phones and agricultural markets in Niger. *American Economic Journal: Applied Economics* 2(3):46–59.

Aquino S (2025) Frubana, startup que captou us$ 271 milhões, encerra operações no brasil e busca rentabilidade em outros mercados. URL https://tinyurl.com/35kzt9ze.

Archetti C, Bertazzi L (2021) Recent challenges in routing and inventory routing: E-commerce and last-mile delivery. *Networks* 77(2):255–268.

Archetti C, Speranza MG (2014) A survey on matheuristics for routing problems. *EURO Journal on Computational Optimization* 2(4):223–246.

Ardestani-Jaafari A, Delage E (2018) The value of flexibility in robust location–transportation problems. *Transp. Sci.* 52(1):189–209.

Balas E (1989) The prize collecting traveling salesman problem. *Networks* 19(6):621–636.

Baldacci R, Mingozzi A, Roberti R (2011) New route relaxation and pricing strategies for the vehicle routing problem. *Oper. Res.* 59(5):1269–1283.

Banerjee D, Erera AL, Toriello A (2025) Pricing and demand management for integrated same-day and next-day delivery systems. *Transp. Sci.* 59(2):279–300.

Belenguer JM, Benavent E, Martínez A, Prins C, Prodhon C, Villegas JG (2016) A branch-and-cut algorithm for the single truck and trailer routing problem with satellite depots. *Transp. Sci.* 50(2):735–749.

Bergquist LF, Dinerstein M (2020) Competition and entry in agricultural markets: Experimental evidence from kenya. *American Economic Review* 110(12):3705–3747.

Bergquist LF, McIntosh C, Startz M (2024) Search costs, intermediation, and trade: Experimental evidence from ugandan agricultural markets. NBER Working Paper 33221, National Bureau of Economic Research.

Bertsimas D, Iancu DA, Katz D (2013) A new local search algorithm for binary optimization. *INFORMS Journal on Computing* 25(2):208–221.

Bian Z, Liu X (2019) Mechanism design for first-mile ridesharing based on personalized requirements part i: Theoretical analysis in generalized scenarios. *Transportation Research Part B: Methodological* 120:147–171.

Bilbao JM (2012) *Cooperative games on combinatorial structures*, volume 26 (Springer Science & Business Media).

Bimpikis K, Candogan O, Saban D (2019) Spatial pricing in ride-sharing networks. *Oper. Res.* 67(3):744–769.

Casaburi L, Reed T (2020) Interlinked transactions and competition: Experimental evidence from cocoa markets, working paper.

Dong L (2021) Toward resilient agriculture value chains: Challenges and opportunities. *Production and Operations Management* 30(3):666–675.

Dror M (1994) Note on the complexity of the shortest path models for column generation in VRPTW. *Oper. Res.* 42(5):977–978.

Engevall S, Göthe-Lundgren M, Värbrand P (2004) The heterogeneous vehicle-routing game. *Transp. Sci.* 38(1):71–85.

EU (2023) Regulation 2023/1115. URL https://tinyurl.com/3tuvhvfx.

EU (2024) Directive of the European Parliament and of the Council on Corporate Sustainability Due Diligence and amending Directive (EU) 2019/1937 and Regulation (EU) 2023/2859. misc.

Fafchamps M, Minten B (2012) Impact of sms-based agricultural information on indian farmers. *The World Bank Economic Review* 26(3):383–414, URL http://dx.doi.org/10.1093/wber/lhr056.

Ferreira K, Goh J, Valavi E (2017) Intermediation in the supply of agricultural products in developing economies. Working paper, Harvard Business School Research Paper Series, sSRN 3047520.

Fischetti M, Lodi A (2003) Local branching. *Mathematical Programming* 98(1–3):23–47.

Fournier N, Guillin A (2015) On the rate of convergence in wasserstein distance of the empirical measure. *Probability theory and related fields* 162(3):707–738.

Gounaris CE, Wiesemann W, Floudas CA (2013) The robust capacitated vehicle routing problem under demand uncertainty. *Oper. Res.* 61(3):677–693.

Gu GY (2024) Technology and disintermediation in online marketplaces. *Man. Sci.* 70(11):7868–7891.

Guajardo M, Rönnqvist M (2016) A review on cost allocation methods in collaborative transportation. *International transactions in operational research* 23(3):371–392.

Gur Y, Iancu D, Warnes X (2021) Value loss in allocation systems with provider guarantees. *Man. Sci.* 67(6):3757–3784.

Hagiu A, Wright J (2024) Marketplace leakage. *Man. Sci.* 70(3):1529–1553.

Hojny C, Gally T, Habeck O, Lüthen H, Matter F, Pfetsch ME, Schmitt A (2020) Knapsack polytopes: a survey. *Annals of Oper. Res.* 292:469–517.

Hoogeboom M, Adulyasak Y, Dullaert W, Jaillet P (2021) The robust vehicle routing problem with time window assignments. *Transp. Sci.* 55(2):395–413.

Hu S, Dessouky MM, Uhan NA, Vayanos P (2021) Cost-sharing mechanism design for ride-sharing. *Transportation Research Part B: Methodological* 150:410–434.

International Finance Correspondent (2025) The rise & fall of eFishery. URL https://tinyurl.com/bdhhu7ks.

Kalkanci B, Rahmani M, Toktay LB (2019) The role of inclusive innovation in promoting social sustainability. *Production and Operations Management* 28(12):2960–2982.

Kamal P (2024) Dehaat reports gross revenue growth in FY24, losses shrink. Accessed 2025-08-09.

Labbé M, Laporte G, Mercure H (1991) Capacitated vehicle routing on trees. *Oper. Res.* 39(4):616–622.

Levi R, Rajan M, Singhvi S, Zheng Y (2020) The impact of unifying agricultural wholesale markets on prices and farmers' profitability. *Proceedings of the National Academy of Sciences* 117(5):2366–2371.

Levi R, Rajan M, Singhvi S, Zheng Y (2024) Improving farmers' income on online agri-platforms: Evidence from the field. *Manufacturing & Service Operations Management* .

Lovász L (1983) Submodular functions and convexity. *Mathematical Programming The State of the Art* 235–257.

Mehta A, Roughgarden T, Sundararajan M (2007) Beyond Moulin mechanisms. *Proceedings of the 8th ACM Conference on Electronic Commerce*, 1–10.

Mohajerin Esfahani P, Kuhn D (2018) Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming* 171(1):115–166.

Özener OÖ, Ergun Ö, Savelsbergh M (2013) Allocating cost of service to customers in inventory routing. *Oper. Res.* 61(1):112–125.

Ross GT, Soland RM (1975) A branch and bound algorithm for the generalized assignment problem. *Mathematical programming* 8(1):91–103.

Schrijver A, et al. (2003) *Combinatorial optimization: polyhedra and efficiency*, volume 24 (Springer).

Sekar S, Siddiq A (2023) Platform disintermediation: Information effects and pricing remedies, SSRN 4378501.

Shi Y, de Zegher JF, Lo I (2023) Two-sided benefits of price transparency in smallholder supply chains, SSRN 4052928.

Sodhi MS, Tang CS (2019) Research opportunities in supply chain transparency. *Prod. and Oper. Management* 28(12):2946–2959.

Subramanyam A, Repoussis PP, Gounaris CE (2020) Robust optimization of a broad class of heterogeneous vehicle routing problems under demand uncertainty. *INFORMS Journal on Computing* 32(3):661–681.

Svitkina Z, Tardos É (2010) Facility location with hierarchical facility costs. *ACM Trans. on Algorithms* 6(2):1–22.

VC4A (2025) Zaidi Technologies. `https://vc4a.com/ventures/zaidi-technologies/`.

Warnes X, de Zegher JF, Iancu D, Plambeck E (2025) Area conditions and positive incentives: Engaging local communities to protect forests, SSRN 4609761.

Waßmuth K, Köhler C, Agatz N, Fleischmann M (2023) Demand management for attended home delivery—a literature review. *European Journal of Operational Research* 311(3):801–815.

Zhou J, Fan X, Chen YJ, Tang CS (2021) Information provision and farmer welfare in developing economies. *Manufacturing & Service Operations Management* 23(1):230–245.

Electronic Companion for

Stable and Profitable Trading Platforms for Smallholder Commodity Supply Chains

Appendix A: Proofs for Results in §2

Proof of Proposition 1. A distribution $\mathbb{P}_t \in \Delta_{\mathcal{F}}$ belongs to $\mathcal{P}_t^{\mathbb{W}}$ if and only if there exists a joint distribution $\{\gamma_{u,v}\}_{u,v\in\{0,1\}^{\mathcal{F}}}$ with marginals $\hat{\mathbb{P}}_t$ and $\mathbb{P}_t$ such that $\sum_{u,v\in\{0,1\}^{\mathcal{F}}} \gamma_{u,v}||u-v||_{1,w} \le \epsilon_t$. Moreover, since $\sum_{u\in\{0,1\}^{\mathcal{F}}} \gamma_{u,v} = \hat{\mathbb{P}}_t(v)$, then $\gamma_{u,v} = 0$ whenever $\hat{\mathbb{P}}_t(v) = 0$. We can write $\hat{\pi}_t$ as:

$$\hat{\pi}_t(r) = \sup_{\gamma\ge 0} \sum_{u\in\{0,1\}^{\mathcal{F}}} \sum_{v\in\{0,1\}^{\mathcal{F}}:\hat{\mathbb{P}}_t(v)>0} \gamma_{u,v}\left(\max_{d_t\in\mathcal{I}_t: d_t\le u}\left(\sum_{f\in\mathcal{F}}(p\cdot q_f - r_f)d_{tf} - c_t(d_t)\right)\right) \tag{EC1a}$$

$$\text{s.t.} \sum_{u\in\{0,1\}^{\mathcal{F}}} \gamma_{u,v} = \hat{\mathbb{P}}_t(v) \tag{EC1b}$$

$$\sum_{u,v\in\{0,1\}^{\mathcal{F}}} \gamma_{u,v}||u-v||_{1,w} \le \epsilon_t \tag{EC1c}$$

$$\gamma_{u,v} = 0 \quad \forall v \notin \text{supp}(\hat{\mathbb{P}}_t) \tag{EC1d}$$

We first claim that without loss of optimality, we can restrict γ in the problem above to satisfy $\gamma_{u,v} = 0$ for any u with $u_f < v_f$ for some f. Consider a feasible γ in (EC1) that satisfies $\gamma_{u,v} > 0$ for some $u, v \in \{0,1\}^{\mathcal{F}}$ with $u_f < v_f$. Define $\gamma' = \gamma - \gamma_{u,v}\cdot\mathbf{1}_{u,v} + \gamma_{u,v}\cdot\mathbf{1}_{u\vee v,v}$. (That is, γ' transfers all the mass that γ assigns to component (u,v) onto component $(u\vee v, v)$ and is the same as γ otherwise.) γ' is clearly feasible in eqs. (EC1b) and (EC1d). To see that γ' is also feasible in (EC1c), it suffices to note that $||(u\vee v)-v||_{1,w} \le ||u-v||_{1,w}$. To show that γ' yields a larger objective than γ, note that the optimal value of the inner maximization problem over d_t in (EC1a) is increasing in u, so the coefficient that multiplies $\gamma_{u,v}$ in (EC1a) is smaller than the coefficient that multiplies $\gamma_{(u\vee v),v}$. Then, because γ' and γ allocate the same total mass on (u,v) and $(u\vee v, v)$, but γ' allocates more on the latter, we conclude that γ' yields a (weakly) larger objective than γ.

In view of this result, and recognizing that $||u\vee v - v||_{1,w} = \sum_{f\in\mathcal{F}} w_f(\mathbf{1}(v_f = 0)\cdot u_f)$ holds for any $u, v \in \{0,1\}^{\mathcal{F}}$, we can rewrite (EC1c) without loss of optimality, as:

$$\sum_{u,v\in\{0,1\}^{\mathcal{F}}} \gamma_{u,v}\left(\sum_{f\in\mathcal{F}} w_f\cdot\mathbf{1}(v_f = 0)\cdot u_f\right) \le \epsilon_t. \tag{EC2}$$

Consider any $\bar{u}\in\{0,1\}^{\mathcal{F}}$, let $d^\star(\bar{u})$ denote the optimal value in the inner maximization problem in (EC1a), and define γ' such that $\gamma'_{\bar{u},v} = 0$ for every $v\in\{0,1\}^n$, $\gamma'_{d^\star(\bar{u}),v} = \gamma_{d^\star(\bar{u}),v} + \gamma_{\bar{u},v}$, and $\gamma'_{u,v} = \gamma_{u,v}$ otherwise. (That is, γ' transfers all the mass that γ assigns to component $(\bar{u},v)$ onto component $(d^\star(\bar{u}),v)$ for every v, and equals γ for all other components.) We claim that γ' is feasible in (EC1) and yields the same objective value as γ. Replacing $\bar{u}$ with $d^\star(\bar{u})$ yields the same optimal value in that inner maximization problem in (EC1a), which implies that γ' yields the same value as γ in (EC1a). Moreover, γ' is feasible in (EC1b),(EC1d) and also in (EC2) (because $d^\star(\bar{u})\le\bar{u}$).

This implies that the following problem with variables $\{\gamma_{d,v}\}_{d\in\mathcal{I}_t, v\in\text{supp}(\hat{\mathbb{P}}_t)}$ has the same optimal value as (EC1):

$$\sup_{\gamma\ge 0} \sum_{d\in\mathcal{I}_t} \sum_{v\in\text{supp}(\hat{\mathbb{P}}_t)} \gamma_{d,v}\left(\sum_{f\in\mathcal{F}}(p\cdot q_f - r_f)d_f - c_t(d)\right) \tag{EC3a}$$

$$\text{s.t.} \sum_{d \in \mathcal{I}_t} \gamma_{d,v} = \hat{\mathbb{P}}_t(v) \quad \forall v \in \text{supp}(\hat{\mathbb{P}}_t) \tag{EC3b}$$

$$\sum_{d \in \mathcal{I}_t} \sum_{v \in \text{supp}(\hat{\mathbb{P}}_t)} \gamma_{d,v} \left(\sum_{f \in \mathcal{F}} w_f \cdot \mathbf{1}(v_f = 0) \cdot d_f \right) \leq \epsilon_t . \tag{EC3c}$$

We claim that the linear program (LP) above has the same optimal value as (8). The LP is feasible; for instance, γ obtained as $\gamma_{\bar{d}(v),v} = \hat{\mathbb{P}}(v)$ for some choice of $\bar{d}(v) \leq v$ with $\bar{d}(v) \in \mathcal{I}_t$ and $\gamma_{d',v} = 0$ for any $d' \neq \bar{d}(v)$ is feasible. (Note that (EC3c) is feasible because the left-hand-side in the constraint is zero, as $v_f = 0$ implies $\bar{d}(v)_f = 0$.) Its objective is upper bounded by $\sum_{f \in \mathcal{F}} p \cdot q_f$, so strong duality holds. The claim follows because the dual of this LP is exactly (8).

Finally, for each v with $\hat{\mathbb{P}}_t(v) > 0$, one can verify if (8b) holds by solving:

$$\max_{d_t \in \mathcal{I}_t} \left(\sum_{f \in \mathcal{F}} \left(p \cdot q_f - r_f - \eta_t \cdot w_f \cdot \mathbf{1}(v_f = 0) \right) d_{tf} - c_t(d_t) \right),$$

which is an instance of oracle (5). □

Proof of Proposition 2 Let $u^\star$ be a feasible, minimum-cost matching, i.e., an optimal solution to the problem:

$$\min_{u \in \{0,1\}^{\mathcal{T} \times \mathcal{F}}} \sum_{t \in \mathcal{T}} c_t(u_t) \quad \text{s.t.} \sum_{t \in \mathcal{T}} u_{tf} = 1 \ \forall f \in \mathcal{F}, \quad u_t \in \mathcal{I}_t \ \forall t \in \mathcal{T}. \tag{EC4}$$

Because $\delta_{u_t^\star} \in \mathcal{P}_t$ by the standing assumption in this proposition, intermediary t can deviate with $u_t^\star$ in problem (PFP). It is simple to verify from (2) that whenever $s_t \neq 0$ in (PFP), it is optimal to set:

$$z_t = \max(c_t(s_t), c_t(s_t) + \hat{\pi}_t(r)) \geq \max\left(c_t(s_t),\ c_t(s_t) + \sum_{f \in \mathcal{F}} p \cdot q_f \cdot u^\star_{tf} - c_t(u_t^\star) \right) \geq c_t(s_t) + \sum_{f \in \mathcal{F}} p \cdot q_f \cdot u^\star_{tf} - c_t(u_t^\star).$$

Therefore, the platform's optimal profit satisfies:

$$\Pi^\star \leq \sum_{f \in \mathcal{F}} (p \cdot q_f - r_f) - \sum_{t \in \mathcal{T}} \left(c_t(s_t) + \left(\sum_{f \in \mathcal{F}} p \cdot q_f u^\star_{tf} - c_t(u_t^\star) \right) \right) = \sum_{t \in \mathcal{T}} c_t(u_t^\star) - \sum_{t \in \mathcal{T}} c_t(s_t) \leq 0,$$

where we used the fact that $r \geq 0$ and that s is feasible for (EC4). This concludes the proof. □

Proof of Proposition 3 From the formulation of $(\text{PFP})_2$ in (9), note that fixing $y_t = 1 - x_{t0}$ makes constraints (9e) depend only on y and π. So problem $(\text{PFP})_2$ splits into separable optimization problems over x and over (r, π), respectively, and the problem of finding an optimal x exactly corresponds to (10). □

Appendix B: Case study supplement

B.1. Data collection

For the case study, we track 14 intermediaries with GPS devices mounted on their trucks, over a period of three months. Some label the nature of their truck stops, and if the stop is for fruit collection, harvested quantities are also labeled. Missing labels are completed using a Machine Learning (ML) pipeline. This pipeline first identifies plantation stops using two filters: (i) a logistic regression (AUC = 0.84) that distinguishes fruit collection from short, non-collection stops based on duration, and (ii) the observed biweekly regularity of farmer–intermediary transactions, as reported by intermediaries, which removes stops that do not follow this pattern. The pipeline then estimates harvested quantities using an L1-penalized linear regression that uses stop duration as a feature. The resulting dataset predicts the locations and harvest volumes of 325 farmers.

B.2. Estimation of transportation costs

To model routing costs, we use the road network estimated by Stienen et al. (2024), which combines satellite-labeled roads with GPS truck traces from our study region. For computational tractability in the (PFP), we approximate the road infrastructure with a Steiner tree spanning all farmer locations and the processing mill, constructed using the KMB algorithm (Kou et al. 1981). Routing costs $c_t(s)$ are then defined as the length of the minimum TSP tour on this tree graph G that departs from the mill, visits all farmers with $s_f = 1$, and returns to the mill. Edge costs in G are calibrated separately for paved and unpaved roads. For paved segments, we estimate 2,625 Rp/km for 9-ton trucks, covering both fuel and maintenance. Fuel costs are derived from Pertamina diesel prices (2018–2021), fuel efficiency ranges from the U.S. Energy Information Administration (US Energy Information Administration 2022), the Indonesian Truckers Club Telegram Lecture (truckmagz 2020), and survey data on Indonesian truck fleets (Yngwie Yudhistira W 2019), with values validated against field observations. Maintenance costs are based on reports from intermediaries. For unpaved roads, where direct measurement is difficult, we adopt a revealed-preference approach: using roughly 1,500 observed routes, we estimate a logistic discrete choice model (cf. Train 2009) that compares realized routes with shortest-path alternatives. The implied cost is 10,320 Rp/km for unpaved segments.

B.3. Generation of Synthetic Instances

To generate synthetic instances, we randomize key parameters using empirical distributions from observed data. Each intermediary's fixed cost σ_t is drawn uniformly from $[\$50, \$140]$, matching the daily rental cost of a truck observed in the field data. Farmer harvests are drawn uniformly from $[0.1, 4.4]$ tons, consistent with observed quantities. Trading relationships are modeled by sampling each nominal distribution $\{\hat{\mathbb{P}}_t\}_{t\in\mathcal{T}}$ uniformly from the 14 observed cases, and when evaluation heterogeneity in ϵ_t, data is constructed sampling it uniformly from $[0, 6]$.

B.4. Intermediary Truck Utilization and Verifying the Toy Model Assumptions

Figure EC1 depicts truck utilization in our historical data. Note that the vast majority of intermediaries do not fully utilize their trucks, so there is ample supply of trucking capacity, which motivates our modeling assumption in §2.

We use this and other data to also verify Assumption 4 in the stylized model from §5. We define intermediaries as type h if they collect at least 5 tons, and as type ℓ otherwise. Figure EC1 shows few trucks are of type h. Conditions (a)–(c) hold by definition. Condition (d) holds under this definition of types and remains valid if the 5-ton threshold is increased. For condition (e), we sample one of the 14 days, select one h- and one ℓ-type intermediary, and draw fixed costs $\sigma_{\ell,h}$ following the procedure in §B.3; we see that condition (e) holds in over 95% of the draws. Condition (f) holds for all σ_ℓ in the ranges of §B.3, even when taking c_f conservatively as a mill–farmer round trip cost. Condition (g) holds when n_ℓ equals the average fruit collected by ℓ-type intermediaries under the status quo.

B.5. Business Constraints and Interpretability

Platforms may impose additional constraints to enhance fairness, transparency, or interpretability, and our algorithmic framework can accommodate several of these. Payments, for instance, can follow structured, interpretable rules such as $r_f = r_{0f} + r_1 q_f$, where r_{0f} depends only on observable characteristics (e.g., location) and $r_1 > 0$ is a uniform per-ton price. Platforms could also prioritize specific intermediaries by requiring that t_2 is matched only if t_1 is (i.e., $y_{t_1} \geq y_{t_2}$), or introducing minimum-income guarantees through lower bounds on payments. Importantly, such requirements are expressible linearly, and convex constraints on (r, π) that do not involve matching variables leave our scheme unchanged.

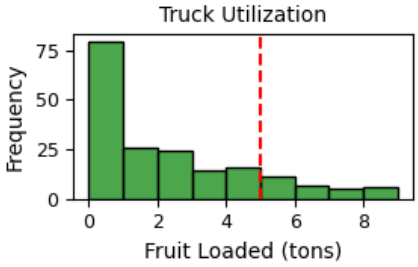

Figure EC1 **Truck utilization for 14 intermediaries who own 9-ton trucks, observed over a period of 14 days.**

For our case study, we evaluate two interpretability constraints. The first is a structured pricing rule in which farmers receive a uniform per-ton price plus two fixed bonuses, one for proximity to a mill and another for low paved-road travel, with thresholds chosen via cross-validation to maximize profit. The second is a matching domination rule: if intermediary t_1 is more cost-efficient and has higher deviation potential than intermediary t_2—i.e., $c_{t_1}(s) \le c_{t_2}(s)$ for all $s \in \mathcal{I}_{t_1}$ and $\rho_{t_1} > \rho_{t_2}$—then $\pi_{t_1} \ge \pi_{t_2}$ is enforced. As shown in Figure EC2, matching domination has negligible effect on profits, whereas structured pricing reduces profit. This confirms the results in Figure 5 and Proposition 5, which demonstrate that – even after controlling for costs – optimal farmer payments depend on intermediaries' deviation potential, highlighting the tension created when pricing is limited to location-based features alone.

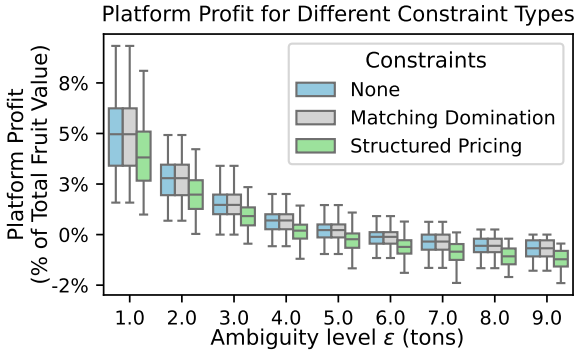

Figure EC2 **Impact on platform profit from including business constraints for *structured pricing* or *domination matching*.**

Appendix C: Proofs for Results in §5

Proof of Lemma 1. **Part (i).** To prove that replacing $\mathcal{P}_t^{\mathbb{W}}$ with $\mathcal{P}_t^1$ from (16a) is without loss, recall the expression of the ambiguity set $\mathcal{P}_t^{\mathbb{W}}$ and that the Wasserstein distance $\mathbb{W}(\mathbb{P}_t, \hat{\mathbb{P}}_t)$ is given by

$$\mathbb{W}(\mathbb{P}_t, \hat{\mathbb{P}}_t) = \inf_{\gamma_{a,b} \ge 0 : a,b \in \{0,1\}^{\mathcal{F}}} \sum_{a,b \in \{0,1\}^{\mathcal{F}}} \gamma_{a,b} ||a - b||_1 \tag{EC5a}$$

$$\text{s.t.} \quad \sum_{b \in \{0,1\}^{\mathcal{F}}} \gamma_{a,b} = \mathbf{1}(a = \mathbf{1}_{\mathcal{H}_t}) \quad \forall a \in \{0,1\}^{\mathcal{F}} \tag{EC5b}$$

$$\sum_{a \in \{0,1\}^{\mathcal{F}}} \gamma_{a,b} = \mathbb{P}_t[b], \quad \forall b \in \{0,1\}^{\mathcal{F}} \tag{EC5c}$$

Note that (EC5b) implies $\gamma_{a,b} = 0$ for all $a \neq \mathbf{1}_{\mathcal{H}_t}$ and all $b \in \{0,1\}^{\mathcal{F}}$. Using this and the notation $\eta_b := \gamma_{\mathbf{1}_{\mathcal{H}_t},b}$ allows rewriting problem (EC5) as:

$$\mathbb{W}(\mathbb{P}_t, \hat{\mathbb{P}}_t) = \inf_{\eta_b \geq 0 : b \in \{0,1\}^{\mathcal{F}}} \sum_{b \in \{0,1\}^{\mathcal{F}}} \eta_b ||\mathbf{1}_{\mathcal{H}_t} - b||_1 \quad \text{s.t.} \quad \eta_b = \mathbb{P}_t[b], \tag{EC6a}$$

which implies that $\mathbb{W}(\mathbb{P}_t, \hat{\mathbb{P}}_t) = \mathbb{E}_{u_t \sim \mathbb{P}_t} ||\mathbf{1}_{\mathcal{H}_t} - u_t||_1$. Rewriting the value of this norm yields (16a).

To prove that using $\mathcal{P}_t^2$ from (16b) is without loss of optimality, consider the proof of Proposition 1. Because our model here is a special instance of the model in Proposition 1, (EC2) also holds here. Note that if we input $w_f = 1$ into that formula and use the fact that $\gamma_{u,v} = 1$ only if $v = \mathbf{1}_{\mathcal{H}_t}$ (and otherwise equal to zero), we obtain the result.

Part (ii). Take any distribution $\mathbb{P}_t$ that is optimal for:

$$\sup_{\mathbb{P}_t \in \mathcal{P}_t^2} \left(\mathbb{E}_{u_t \sim \mathbb{P}_t} \left[\max_{d_t \leq u_t : d_t \in \mathcal{I}_t} (p - r - c)^\top d_t \right] \right) - \sigma_t. \tag{EC7}$$

We first show that the optimal value in problem (17) upper bounds the optimal value of (EC7). Take a random vector $u_t \sim \mathbb{P}_t$ and let $d_t^\star(u_t)$ be any vector that solves the inner maximization problem for u_t. It is simple to verify that $d_{tf}^\star(u_t) = 0$ if $p - r_f - c_f < 0$. Therefore,

$$d_t^\star(u_t) \in \operatorname*{arg\,max}_{y_t \leq u_t : y_t \in \mathcal{I}_t} \left(\sum_{f \in \mathcal{F}} \max(p - r_f - c_f, 0) \cdot y_{tf} - \sigma_t \right). \tag{EC8}$$

Define $\mu_{tf} = \mathbb{P}_t[d_{tf}^\star(u_t) = 1]$. Note that:

$$\mathbb{E}_{u_t \sim \mathbb{P}_t} \left[\max_{d_t \leq u_t : d_t \in \mathcal{I}_t} \left(\sum_{f \in \mathcal{F}} (p - r_f - c_f) d_{tf} - \sigma_t \right) \right] = \sum_{f \in \mathcal{F}} \mu_{tf} \cdot \max(p - r_f - c_f, 0). \tag{EC9}$$

Now, the reformulation (16b) of the ambiguity set implies that:

$$\sum_{f \in \mathcal{H}_t^c} \mu_{tf} = \mathbb{E}_{u_t \sim \mathbb{P}_t} \left[\sum_{f \in \mathcal{H}_t^c} d_{tf}^\star(u_t) \right] \leq \mathbb{E}_{u_t \sim \mathbb{P}_t} \left[\sum_{f \in \mathcal{H}_t^c} u_{tf} \right] \leq \epsilon_t, \tag{EC10}$$

where the second inequality follows because $d_t^\star(u_t) \leq u_t$ holds almost surely. Finally, because $d_t^\star(u_t) \in \mathcal{I}_t$, we have that $\sum_{f \in \mathcal{F}} d_{tf}^\star(u_t) \leq K$ implies that $\sum_{f \in \mathcal{F}} \mu_{tf} \leq K$, which means that μ_t is feasible for (17) and yields an objective value at most equal to that of (17).

Now, we prove that the optimal value of (17) is a lower bound on that of (EC7), which will finish the proof. Observe that the feasibility set of (17) is the intersection of two matroids, therefore, by the intersection matroid theorem, specifically, Theorem 41.12 of Schrijver et al. (2003), any μ_t that is optimal for (17) can be assumed to be integral. This is equivalent to $\mu_t \in \{0,1\}^{\mathcal{F}}$. Now, consider $\delta_{\mu_t} \in \Delta_{\mathcal{F}}$ as the distribution that places probability one on μ_t. To verify that $\delta_{\mu_t} \in \mathcal{P}_t^2$, note that: $\mathbb{E}_{u_t \sim \delta_{\mu_t}} \left[\sum_{f \in \mathcal{H}_t^c} u_{tf} \right] = \sum_{f \in \mathcal{H}_t^c} \mu_{tf} \leq \epsilon_t$, which shows that $\delta_{\mu_t} \in \mathcal{P}_t^2$. This means that the optimal value in (EC7) is at least:

$$\max_{d_t \leq \mu_t : d_t \in \mathcal{I}_t} \left(\sum_{f \in \mathcal{F}} (p - r_f - c_f) d_{tf} - \sigma_t \right) \leq \sum_{f \in \mathcal{F}} \max(p - r_f - c_f, 0) \mu_{tf} - \sigma_t, \tag{EC11}$$

which is the desired result.

Finally, notice that when solving (17), farmers to deviate with are selected greedily in descending order of the margin $p - r_f - c_f$ and avoiding them when such margin is negative, with the constraint that at most ϵ_t farmers are selected from $\mathcal{H}_t^c$ and at most K farmers are selected in total.

Part (iii). Re-express problem (14) with decisions v_f instead of r_f. Because the objective is decreasing in π_t, the constraints $\pi \geq \hat{\pi}_t(r)$ and $\pi \geq 0$ together with the fact that $\hat{\pi}_t(r) = \tilde{\pi}_t(v)$ (by Lemma 1) imply that it is optimal to set $\pi_t = \max(\tilde{\pi}_t(v), 0)$ for all matched intermediaries $t \in T$. And because (14d) requires $\pi_t = 0$ for all unmatched intermediaries $t \in T^c$, we can replace this with $0 = \tilde{\pi}_t(v)$ and further relax this into constraint (18b), without affecting optimality (the objective does not depend on $t \in T^c$). The requirement $r_f \geq 0$ is equivalent to $v_f \leq p - c_f$. Lastly, it is simple to verify from Lemma 1 that the deviation profit satisfies $\tilde{\pi}_t(v) = \tilde{\pi}_t\big(\max(0, v)\big)$ for any v, and because the platform's objective is increasing in v_f for all $f \in \mathcal{F}$, we can constrain $v_f \geq 0$ without loss of optimality. □

Proof of Proposition 4 The proof relies on Lemma 2 from §C.1. By Lemma 2(iv)-(v), the optimal value in $(\text{S-PFP})_2$ equals the value of (EC21) for any $(x_h, \bar{v}) \in X^\star$. To determine the optimal platform profit and whether optimal cost-minimizing matches T exist, it suffices to characterize the objective in (EC21), denoted with $\Pi(x_h, \bar{v})$, over choices $(x_h, \bar{v})$ at the corners of $X^\star$ (due to $X^\star$ being a face of the polytope $[0, |\mathcal{T}_h|] \times [e_h, e_\ell]$ Lemma 2(vi)). Following the notation in the proof of Lemma 2-(v), let $A_\theta = \min(n_\theta + \epsilon_\theta, K)$, $x_\theta = |T \cap \mathcal{T}_\theta|$, $\forall \theta \in \{\ell, h\}$, $\delta_h = \frac{\max(0, n_h + \epsilon_h - K)}{\min(n_h, K - \epsilon_h)}$, $e_\ell := \frac{\sigma_\ell}{n_\ell + \epsilon_\ell}$, and $e_h := \frac{\sigma_h}{\min(n_h + \epsilon_h, K)}$. The corners of $X^\star$ satisfy $\bar{v} \in \{e_\ell, e_h\}$, so we can express $\Pi(x_h, \bar{v})$ for these two cases:

$$\Pi(x_h, e_h) = |\mathcal{F}| e_h - (|\mathcal{F}|/K - x_h)\sigma_\ell - x_h \sigma_h \tag{EC12}$$

$$\Pi(x_h, e_\ell) = |\mathcal{F}| e_\ell - \big(|\mathcal{T}_h| + \delta_h(|\mathcal{T}_h| - x_h)\big)(e_\ell - e_h)A_h - (|\mathcal{F}|/K - x_h)\sigma_\ell - x_h \sigma_h. \tag{EC13}$$

Subsequently, we denote: $c_h^\star = \arg\min_{x_h \in [0, |\mathcal{T}_h|]}((|\mathcal{F}|/K - x_h)\sigma_\ell + x_h \sigma_h)$, so that a choice $x_h = c_h^\star$ yields a matching that minimizes total costs. Note that $\sigma_h \neq \sigma_\ell$ implies $c_h^\star \in \{0, |\mathcal{T}_h|\}$.

We characterize the optimal $(x_h, \bar{v})$ and when this yields a matching that minimizes transportation costs.

Case I: $n_h + \epsilon_h \leq K$. We prove that $(c_h^\star, e_\ell)$ is the optimal solution. Note that in this case $\delta_h = 0$, and so both (EC12) and (EC13) are maximized at $x_h = c_h^\star$. Moreover, note that:

$$\Pi(c_h^\star, e_\ell) - \Pi(c_h^\star, e_h) = (|\mathcal{F}| - |\mathcal{T}_h|(n_h + \epsilon_h))(e_\ell - e_h) > 0, \tag{EC14}$$

so the optimal profit is $\Pi(c_h^\star, e_\ell)$ and the optimal matching is a cost-minimizing one.

Case II: $\sigma_h < \sigma_\ell$ and $n_h + \epsilon_h > K$. We prove that $\Pi(|\mathcal{T}_h|, e_\ell) = \Pi(c_h^\star, e_\ell)$ is the unique optimal solution. Note that $n_h = |\mathcal{T}_h| = c_h^\star$ maximizes (EC12) because $\sigma_h < \sigma_\ell$, and also maximizes (EC13) because (EC13) increases strictly in x_h. Then,

$$\Pi(|\mathcal{T}_h|, e_\ell) - \Pi(|\mathcal{T}_h|, e_h) = (|\mathcal{F}| - |\mathcal{T}_h|K)(e_\ell - e_h) > 0, \tag{EC15}$$

so the optimal profit is $\Pi(|\mathcal{T}_h|, e_\ell)$ and the optimal matching is a cost-minimizing one.

Note that in Cases I and II, which correspond to Cases (i) and (ii) of Proposition 4, the optimal profit is given by:

$$\begin{aligned}
\Pi(c_h^\star, e_\ell) &= |\mathcal{F}|\frac{\sigma_\ell}{A_\ell} - |\mathcal{T}_h|A_h\left(\frac{\sigma_\ell}{A_\ell} - \frac{\sigma_h}{A_h}\right) - \left(\frac{|\mathcal{F}|}{K} - c_h^\star\right)\sigma_\ell - c_h^\star \sigma_h \\
&= (n_\ell|\mathcal{T}_\ell| + n_h|\mathcal{T}_h|)\frac{\sigma_\ell}{A_\ell} - |\mathcal{T}_h|A_h\frac{\sigma_\ell}{A_\ell} + |\mathcal{T}_h|\sigma_h - \min_{T \subseteq \mathcal{T}: |T| \geq |\mathcal{F}|/K} \sum_{t \in T} \sigma_t \\
&= \Delta C - \frac{\epsilon_\ell|\mathcal{T}_\ell| + \min(\epsilon_h, K - n_h)|\mathcal{T}_h|}{n_\ell + \epsilon_\ell}\sigma_\ell := A,
\end{aligned}$$

where $\Delta C := \sum_{t\in\mathcal{T}} \sigma_t - \min_{\substack{T\subseteq\mathcal{T}\\ |T|\geq|\mathcal{F}|/K}} \sum_{t\in T}\sigma_t$.

Case III: $\sigma_\ell < \sigma_h$ and $n_h + \epsilon_h > K$. We prove that any of $\Pi(0, e_\ell)$, $\Pi(0, e_h)$ and $\Pi(|\mathcal{T}_h|, e_\ell)$ can be optimal. First, note that $x_h = 0$ maximizes (EC12), so we can discard $\Pi(|\mathcal{T}_h|, e_h)$ as an optimal value. With $A_h = K$, we have:

$$\Pi(0, e_h) = |\mathcal{F}|e_h - \frac{|\mathcal{F}|}{K}\sigma_\ell = A - (|\mathcal{F}| - |\mathcal{T}_h|K)(e_\ell - e_h)$$
$$\Pi(|\mathcal{T}_h|, e_\ell) = |\mathcal{F}|e_\ell - |\mathcal{T}_h|K(e_\ell - e_h) - \left(\frac{|\mathcal{F}|}{K} - |\mathcal{T}_h|\right)\sigma_\ell - |\mathcal{T}_h|\sigma_h = A - |\mathcal{T}_h|(\sigma_h - \sigma_\ell)$$
$$\Pi(0, e_\ell) = |\mathcal{F}|e_\ell - |\mathcal{T}_h|K(e_\ell - e_h) - |\mathcal{T}_h|\delta_h K(e_\ell - e_h) - \frac{|\mathcal{F}|}{K}\sigma_\ell = A - |\mathcal{T}_h|\delta_h K(e_\ell - e_h).$$

Therefore, the optimal matching only matches type ℓ when $\max(\Pi(0, e_h), \Pi(0, e_\ell)) > \Pi(|\mathcal{T}_h|, e_\ell)$, which is equivalent to condition $B_\ell > B_h$, and matches all type h when the reverse equality holds. Moreover $B_\ell - B_h > 0$ if and only if:

$$\min\left(|\mathcal{T}_h|\frac{n_h + \epsilon_h - K}{K - \epsilon_h}K, |\mathcal{F}| - |\mathcal{T}_h|K\right)\left(\frac{\sigma_\ell}{n_\ell + \epsilon_\ell} - \frac{\sigma_h}{K}\right) - |\mathcal{T}_h|(\sigma_h - \sigma_\ell) > 0. \tag{EC16}$$

Notice, that the left hand side is continuous, increasing in ϵ_h and σ_ℓ and decreasing in σ_h and ϵ_ℓ, therefore, there exists a function $\theta_h(\sigma_h, \epsilon_\ell, \sigma_\ell)$ which is increasing in σ_h, ϵ_ℓ, decreasing in σ_ℓ, and such that $\epsilon_h > \theta_h(\sigma_h, \epsilon_\ell, \sigma_\ell)$ if and only if condition (EC16) holds. This shows the profits for parts (iii) and (iv) of the Proposition. □

Proof of Proposition 5. The proof relies on Lemma 2. To prove Part (i), note from (EC23c) that intermediary welfare corresponds to (19) when type h intermediaries are matched, and is zero if intermediaries of type h are unmatched.

To prove (ii), note that farmer welfare plus intermediary welfare must be a constant (as transportation costs and the platform optimal profit do not change). Therefore, the dependencies on ϵ_h and ϵ_ℓ are opposite to those of Part (i).

To prove the bound in (20), note that from (EC23b) we have

$$\begin{aligned}\underline{\mathcal{W}}^{\mathcal{F}} &= \mathbf{1}^\top(p - c) - \max_{(x_h,\bar{v})\in X^\star}\left[\bar{v}x_h n_h + \frac{\sigma_h - \epsilon_h\bar{v}}{\min(n_h, K - \epsilon_h)}(|\mathcal{T}_h| - x_h)n_h + \bar{v}|\mathcal{T}_\ell|n_\ell\right]\\ &\leq \mathbf{1}^\top(p - c) - \frac{\sigma_\ell}{n_\ell + \epsilon_\ell}(|\mathcal{T}_h|n_h + \bar{v}|\mathcal{T}_\ell|n_l) = \sum_{f\in\mathcal{F}}\big(p - c_f - \frac{\sigma_\ell}{n_\ell + \epsilon_\ell}\big),\end{aligned}$$

where we use the bounds $\bar{v} \leq \sigma_\ell/(n_\ell + \epsilon_\ell)$ and $\bar{v} \geq \frac{\sigma_h - \epsilon_h\bar{v}}{\min(n_h, K - \epsilon_h)}$ (which is equivalent to $\bar{v} \geq \sigma_h/\min(K, n_h + \epsilon_h)$).

Finally, note from conditions (EC22c)-(EC22f) that it is always possible to choose $v^t = v^\theta$ for some v^θ and all $t \in \mathcal{T}_\theta$ and all $\theta \in \{\ell, h\}$. Moreover, the choice always would set $v^\ell = \bar{v}$, and so $v^\ell \geq v^h$, proving (21). □

C.1. Supporting Lemmas

Definition 1 (Margin-equalizing). A vector $v \in \mathbb{R}^{\mathcal{T}}$ is said to be *margin-equalizing* if the components corresponding to farmers transacting with the *same* high-type intermediary are equal and the components corresponding to *all* farmers transacting with low-type intermediaries are equal, i.e.,

$$v_f = v_g, \forall\, f, g \in \mathcal{H}_t, \forall\, t \in \mathcal{T}_h \quad \text{and} \quad v_f = v_g, \forall\, f, g \in \mathcal{H}[\mathcal{T}_\ell]. \tag{EC17}$$

Let $\mathcal{S}^{\text{mg-eq}}$ denote the set of all pairs (T, v) such that v is margin-equalizing and define the *margin-equalizing function*:

$$\mathrm{f}^{\text{mg-eq}} : \mathbb{R}^{\mathcal{F}} \to \mathcal{S}^{\text{mg-eq}}, \quad \mathrm{f}^{\text{mg-eq}}(v') = v \text{ where } v_f = \begin{cases}\frac{1}{|\mathcal{H}_t|}\sum_{g\in\mathcal{H}_t} v'_g & \text{if } f \in \mathcal{H}_t \text{ for some } t \in \mathcal{T}_h\\ \frac{1}{|\mathcal{H}[\mathcal{T}_\ell]|}\sum_{g\in\mathcal{H}[\mathcal{T}_\ell]} v'_g & \text{if } f \in \mathcal{H}[\mathcal{T}_\ell].\end{cases} \tag{EC18}$$

Lemma 2 (**Relaxing (S-PFP)$_2$; Structural Results for Relaxation.**). *With $\tilde{\pi}_t(v)$ as defined in* (17)*, we have:*

(i) The optimal value in (S-PFP)$_2$ *is upper bounded by the optimal value in the relaxed problem:*

$$\text{(RS-PFP)} \qquad \max_{T\subseteq\mathcal{T},\, v\in\mathbb{R}^{\mathcal{F}}} \sum_{f\in\mathcal{F}} v_f - \sum_{t\in\mathcal{T}} \max\big(\tilde{\pi}_t(v),0\big) - \sum_{t\in T}\sigma_t \tag{EC19a}$$

$$\text{s.t. } 0 \geq \tilde{\pi}_t(v), \quad \forall t\in T^c\cap\mathcal{T}_h \tag{EC19b}$$

$$|T| \geq |\mathcal{F}|/K. \tag{EC19c}$$

(ii) If (T,v') *is an optimal solution for (*RS-PFP*) and* $v = \mathrm{f}^{\text{mg-eq}}(v')$ *where* $\mathrm{f}^{\text{mg-eq}}$ *is the margin-equalizing function defined in* (EC18)*, then* (T,v) *is also optimal for (*RS-PFP*) and yields the same farmer and intermediary welfare,* $\mathcal{W}^{\mathcal{F}}(v)=\mathcal{W}^{\mathcal{F}}(v')$, $\mathcal{W}^{\mathcal{T}}(v)=\mathcal{W}^{\mathcal{T}}(v')$.

(iii) If (T,v) *is an optimal solution to (*RS-PFP*) where* (T,v) *is margin-equalizing,* $(T,v)\in\mathcal{S}^{\text{mg-eq}}$*, then:*

$$\tilde{\pi}_t(v) = \min(n_t, K-\epsilon_t)v_f + \epsilon_t\bar{v} - \sigma_t, \;\; \textit{for some } f\in\mathcal{H}_t, \tag{EC20}$$

where $\bar{v}=\max_{f\in\mathcal{F}} v_f$ *and* $f\in\mathcal{H}_t$ *is arbitrary (because* $v_f = v_g, \forall f,g\in\mathcal{H}_t$*).*

(iv) Let $X^\star$ *denote the set of maximizers in the optimization problem:*

$$\max_{\substack{x_h\in[0,|\mathcal{T}_h|]\\ \bar{v}\in[e_h,e_\ell]}} \Pi(x_h,\bar{v}) := |\mathcal{F}|\,\bar{v} - |\mathcal{T}_\ell|\big[A_\ell\bar{v}-\sigma_\ell\big]^+ - |\mathcal{T}_h|\big[A_h\bar{v}-\sigma_h\big]^+ - \big(|\mathcal{T}_h|-x_h\big)\delta_h\big[A_h\bar{v}-\sigma_h\big]^+ - (|\mathcal{F}|/K - x_h)\sigma_\ell - x_h\sigma_h, \tag{EC21}$$

where $e_\ell := \frac{\sigma_\ell}{n_\ell+\epsilon_\ell}$, $e_h := \frac{\sigma_h}{\min(n_h+\epsilon_h,K)}$, $A_\theta := \min(n_\theta+\epsilon_\theta, K)$*, and* $\delta_h := \frac{\max(0,n_h+\epsilon_h-K)}{\min(n_h,K-\epsilon_h)}$*. Then, the optimal value of this problem equals the optimal value of* (RS-PFP)*, and the set of optimal solutions* (T,v) *to* (RS-PFP) *with* $v\in\mathcal{S}^{\text{mg-eq}}$ *is:*

$$\mathcal{S}^{\star,\text{mg-eq}} = \Big\{(T,v)\in\mathcal{T}\times\mathbb{R}^{\mathcal{F}} \textit{ such that } \exists v^t\in\mathbb{R}, \forall t\in\mathcal{T}: v_f = v^t, \qquad \forall f\in\mathcal{H}_t, \forall t\in\mathcal{T} \tag{EC22a}$$

$$\exists (x_h,\bar{v})\in X^\star : \bar{v} = \max_{f\in\mathcal{F}} v_f,\; x_h = |T\cap\mathcal{T}_h|, \tag{EC22b}$$

$$v^t = \bar{v} \qquad \forall t\in\mathcal{T}_\ell \tag{EC22c}$$

$$v^t \in \left[\frac{\sigma_h-\epsilon_h\bar{v}}{\min(n_h,K-\epsilon_h)},\bar{v}\right] \qquad \forall t\in T\cap\mathcal{T}_h \textit{ if } n_h+\epsilon_h\leq K \tag{EC22d}$$

$$v^t = \bar{v} \qquad \forall t\in T\cap\mathcal{T}_h \textit{ if } n_h+\epsilon_h > K \tag{EC22e}$$

$$v^t = \frac{\sigma_h-\epsilon_h\bar{v}}{\min(n_h,K-\epsilon_h)} \qquad \forall t\in T^c\cap\mathcal{T}_h\Big\}. \tag{EC22f}$$

(v) (S-PFP)$_2$ *and* (RS-PFP) *have the same optimal value and the same optimal margin-equalizing solutions* $\mathcal{S}^{\star,\text{mg-eq}}$*.*

(vi) The minimum and maximum farmer and intermediary welfare over all optimal solutions to (S-PFP)$_2$ *are given by:*

$$\overline{\mathcal{W}^{\mathcal{F}}} = \mathbf{1}^\top(p-c) - \min_{(x_h,\bar{v})\in X^\star}\left[\left(\mathbf{1}(n_h+\epsilon_h\leq K)\frac{\sigma_h-\epsilon_h\bar{v}}{n_h} + \mathbf{1}(n_h+\epsilon_h>K)\bar{v}\right)x_h n_h + \frac{(\sigma_h-\epsilon_h\bar{v})(|\mathcal{T}_h|-x_h)n_h}{\min(n_h,K-\epsilon_h)} + \bar{v}|\mathcal{T}_\ell|n_\ell\right] \tag{EC23a}$$

$$\underline{\mathcal{W}^{\mathcal{F}}} = \mathbf{1}^\top(p-c) - \max_{(x_h,\bar{v})\in X^\star}\left[\bar{v}x_h n_h + \frac{\sigma_h-\epsilon_h\bar{v}}{\min(n_h,K-\epsilon_h)}(|\mathcal{T}_h|-x_h)n_h + \bar{v}|\mathcal{T}_\ell|n_\ell\right] \tag{EC23b}$$

$$\overline{\mathcal{W}^{\mathcal{T}}} = \max_{(x_h,\bar{v})\in X^\star}\Big[\big(\min(n_t+\epsilon_t,K)\bar{v}-\sigma_h\big)x_h\Big] \tag{EC23c}$$

$$\underline{\mathcal{W}^{\mathcal{T}}} = \min_{(x_h,\bar{v})\in X^\star}\Big[\mathbf{1}(n_h+\epsilon_h>K)(K\bar{v}-\sigma_h)x_h\Big], \tag{EC23d}$$

where $X^\star$ *is defined in (iv). Moreover,* $X^\star$ *is a face of the polytope* $[0,|\mathcal{T}_h|]\times[e_h,e_\ell]$ *and* $\overline{\mathcal{W}^{\mathcal{F}}}, \underline{\mathcal{W}^{\mathcal{F}}}, \overline{\mathcal{W}^{\mathcal{T}}}$ *and* $\underline{\mathcal{W}^{\mathcal{T}}}$ *are achieved at respective optimal solutions* $(x_h^\star,\bar{v}^\star)\in X^\star$ *at a corner of* X^* *(i.e.* $x_h^\star\in\{0,|\mathcal{T}_h|\}$ *and* $\bar{v}^\star\in\{e_\ell,e_h\}$*).*

Proof of Lemma 2 To facilitate following the proof, we repeat the definition $\tilde{\pi}_t(v)$ from (17):

$$\tilde{\pi}_t(v) := -\sigma_t + \sup_{\mu_t \in [0,1]^{\mathcal{F}}} \left\{ \mu_t^\top v \;:\; \sum_{f \in \mathcal{H}_t^c} \mu_{tf} \le \epsilon_t,\ \sum_{f \in \mathcal{F}} \mu_{tf} \le K \right\}, \text{ where } v := p - r - c.$$

Part (i). The proof follows by noting the differences between (EC19) and (18). First, (EC19) adds a term $-\sum_{t \in T^c} \max(\tilde{\pi}_t(v), 0)$ to the objective, but this term is zero for any solution that is feasible in (S-PFP)$_2$ because $\tilde{\pi}_t(v) \le 0$, $\forall t \in T^c$ by (18b). Second, (EC19b) only imposes constraint (18b) for $t \in T^c \cap \mathcal{T}_h$ rather than for $t \in T^c$. Lastly, (EC19) does not include constraints $v_f \le p - c_f$ for all $f \in \mathcal{F}$.

Part (ii). Applying Proposition 7 with the set $S = \mathcal{T}_\ell$, and then again sequentially with set $S = \{t\}$, for every $t \in \mathcal{T}_h$, yields the following inequalities:

$$\sum_{t \in \mathcal{T}_\ell} \max(\tilde{\pi}_t(v'), 0) \le \sum_{t \in \mathcal{T}_\ell} \max(\tilde{\pi}_t(v), 0)$$
$$\max(\tilde{\pi}_t(v'), 0) \le \max(\tilde{\pi}_t(v), 0) \quad \forall t \in \mathcal{T}_h.$$

With $\Pi(T, v)$ and $\Pi(T, v')$ as the objectives of (T, v) and (T, v') in (RS-PFP), respectively, we have:

$$\Pi(T, v) = \sum_{f \in \mathcal{F}} v_f - \sum_{t \in \mathcal{T}} \max(\tilde{\pi}_t(v), 0) - \sum_{t \in T} \sigma_t \le \sum_{f \in \mathcal{F}} v'_f - \sum_{t \in \mathcal{T}} \max(\hat{\pi}_t(v'), 0) - \sum_{t \in T} \sigma_t = \Pi(T, v'),$$

where the inequality follows from the two inequalities above and because $\sum_{f \in \mathcal{F}} v_f = \sum_{f \in \mathcal{F}} v'_f$. Because (T, v) was optimal, equality must then hold throughout. Moreover, because $0 = \max(\tilde{\pi}_t(v), 0) \ge \max(\hat{\pi}_t(v'), 0)$ for $t \in T^c \cap \mathcal{T}_h$, we readily see that (T, v') is feasible in (RS-PFP). Finally, the welfare of farmers is unchanged because $\mathbf{1}^\top v = \mathbf{1}^\top v'$ and because $\Pi(T, v') = \Pi(T, v)$ and the matching T is the same, the welfare of intermediaries is also unchanged.

Part (iii). Because $v \in \mathcal{S}^{\text{mg-eq}}$, we can ease notation by defining $v^\ell = v_f$ for an arbitrary $f \in \mathcal{F}_\ell$ and $v^t = v_f$ for an arbitrary $f \in \mathcal{H}_t$, for every $t \in \mathcal{T}_t$. Choose $t \in \mathcal{T}$ and consider the following (non-exhaustive) cases:

Case I. There are at least ϵ_t farmers in $\mathcal{H}_t^c$ with $v_f = \bar{v}$. A simple interchange argument proves that it is optimal to set $\mu_{tf} = 1$ for ϵ_t farmers $f \in \mathcal{H}_t^c$ with $v_f = \bar{v}$, and set $\mu_{tf} = 1$ for an arbitrary set of $\min(K - \epsilon_t, n_t)$ farmers from $\mathcal{H}_t$ (arbitrary because $v \in \mathcal{S}^{\text{mg-eq}}$). The profit expression matches (EC20).

Case II. $v_t = \bar{v}$ and there are at least K farmers in $\mathcal{F}$ with $v_f = \bar{v}$. A simple interchange argument proves that it is optimal to set $\mu_{ft} = 1$ for all n_t farmers $f \in \mathcal{H}_t$, and set $\mu_{ft} = 1$ for $\min(K - n_t, \epsilon_t)$ farmers in $\mathcal{H}_t^c$ with margin $\bar{v}$. Again, this leads to a profit expression matching (EC20).

Subsequently, we prove that either Case I or Case II must arise by considering all possible cases for t:

(a) $t \in \mathcal{T}_\ell$, $v^\ell = \bar{v}$. Case II arises because Assumption 4(g) ensures that $|\mathcal{H}[\mathcal{T}_\ell]| \ge K$ farmers achieve margin $\bar{v}$.

(b) $t \in \mathcal{T}_\ell$ and $v^t = \bar{v}$ for some $t \in \mathcal{T}_h$. There are at least n_h farmers with margin $\bar{v}$ in $\mathcal{H}[\mathcal{T}_h] \subseteq \mathcal{H}_t^c$. By Assumption 4(b), $n_h \ge \epsilon_\ell$, so Case I arises.

(c) $t \in \mathcal{T}_h$ and $v^\ell = \bar{v}$. Then, note that $\mathcal{H}[\mathcal{T}_\ell] \subseteq \mathcal{H}_t^c$ and $|\mathcal{H}[\mathcal{T}_\ell]| \ge K \ge n_t \ge \epsilon_t$, where the inequalities follow from Assumption 4(g), feasibility of $\mathcal{H}_t$, and Assumption 4(b). Therefore, Case I arises.

(d) $t \in \mathcal{T}_h$ and $v^s = \bar{v}$ for $s \in \mathcal{T}_h$ with $s \ne t$. Because $\mathcal{H}_s \subseteq \mathcal{H}_t^c$ and $|\mathcal{H}_s| = n_h \ge \epsilon_t$ by Assumption 4(b), Case I arises.

(e) $t \in \mathcal{T}_h$, $v^t = \bar{v}$ and $v^t > v^s$ for all $s \in \mathcal{T}$ with $s \ne t$. We will prove that this results in a contradiction. Denote $\bar{v}_2 = \max_{s \in \mathcal{T}: s \ne t} v^s < \bar{v}$ and let t_2 be such that $v^{t_2} = \bar{v}_2$. For any $s \in \mathcal{T}$, we have:

$$\tilde{\pi}_s(v) = \begin{cases} \min(n_s, K - \epsilon_s) v^s + \epsilon_s \bar{v} - \sigma_s & \text{if } s \ne t \\ n_h \bar{v} + \min(\epsilon_h, K - n_h) \bar{v}_2 - \sigma_h & \text{if } s = t, \end{cases} \tag{EC24}$$

where the first case holds because it falls in scenarios (a)-(d) above, and the second case holds because intermediary t would deviate with all n_t farmers in $\mathcal{H}_t$ and an additional $\min(\epsilon_t, K - n_t)$ farmers from $\mathcal{H}_t^c$ with margin $\bar{v}_2$. (There are enough such farmers: if $t_2 \in \mathcal{T}_\ell$, then $\min(\epsilon_t, K - n_t) \leq K \leq |\mathcal{H}[\mathcal{T}_\ell]|$ by Assumption 4(g); if $t_2 \in \mathcal{T}_h$, then $\min(\epsilon_t, K - n_t) \leq \epsilon_t \leq n_{t_2} = n_h$ by Assumption 4(b).) We consider two sub-cases, depending on the value of $\tilde{\pi}_t(v) \leq 0$.

Sub-case (e-1): $\tilde{\pi}_t(v) \leq 0$. We verify from (EC24) that $\tilde{\pi}_s(v) < \tilde{\pi}_t(v), \forall s \in \mathcal{T}_h \setminus \{t\}$. Note that:

$$\begin{aligned}\tilde{\pi}_s(v) &= \min(n_h, K - \epsilon_h)v^s + \epsilon_h \bar{v} - \sigma_h \leq \min(n_h, K - \epsilon_h)\bar{v}_2 + \epsilon_h \bar{v} - \sigma_h \\ &= \tilde{\pi}_t(v) - (n_h - \epsilon_h)(\bar{v} - \bar{v}_2) < \tilde{\pi}_t(v) \quad \text{(because } n_h > \epsilon_h \text{ by Assumption 4(b))}. \end{aligned} \tag{EC25}$$

Consider a high-type intermediary $\tau \neq t \in \mathcal{T}_h$ and define a new margin vector $\hat{v}$ that decreases slightly the margin $\bar{v}$ for all farmers in $\mathcal{H}_t$ and increases slightly *more* the margin for all farmers $f \in \mathcal{H}_\tau$:

$$\hat{v}_f = \begin{cases} v_f - \delta = \bar{v} - \delta & \text{if } f \in \mathcal{H}_t \\ v_f + \delta^+ = v^\tau + \delta^+ & \text{if } f \in \mathcal{H}_\tau \\ v_f & \text{otherwise.} \end{cases}$$

We prove that for sufficiently small $\delta^+ > \delta > 0$, $(T, \hat{v})$ is feasible in (RS-PFP) and yields higher objective than (T, v). For sufficiently small δ^+, δ, we can ensure that $\bar{v} - \delta > \max_{f \notin \mathcal{H}_t} \hat{v}_f$. As before, define $\hat{v}_2^{\max} = \max_{s \in \mathcal{T}: s \neq t} \hat{v}_s < \bar{v} - \delta$. To prove feasibility of $(T, \hat{v})$, note that we have:

$$\begin{aligned}\tilde{\pi}_s(\hat{v}) &= \min(n_s, K - \epsilon_s)v^s + \epsilon_s(\bar{v} - \delta) - \sigma_s = \tilde{\pi}_s(v) - \delta\epsilon_s < \tilde{\pi}_s(v), \ \forall\, s \notin \{t, \tau\}. \\ \tilde{\pi}_t(\hat{v}) &= n_h(\bar{v} - \delta) + \min(\epsilon_h, K - n_h)\hat{v}_2^{\max} - \sigma_h \leq n_h(\bar{v} - \delta) + \min(\epsilon_h, K - n_h)(v_2^{\max} + \delta^+) - \sigma_h \\ &= \tilde{\pi}_t(v) - n_h\delta + (\epsilon_h - \kappa_h)\delta^+, \end{aligned}$$

where in the first inequality we used that $\hat{v}_2^{\max} \leq v_2^{\max} + \delta^+$. Note that the last expression is strictly negative for $\delta^+ = \delta$ (because $\epsilon_h \leq n_h$ by Assumption 4(b)). Because $\tilde{\pi}_t(\hat{v}) \leq 0$ by the standing assumption in this sub-case and $\tilde{\pi}_t(\hat{v})$ is a continuous function of δ, δ^+, we conclude that $\delta^+ > \delta$ exists such that $\tilde{\pi}(\hat{v}) \leq 0$ still holds. Lastly, consider intermediary τ: because $\tilde{\pi}_\tau(v) < \tilde{\pi}_t(v) \leq 0$ by (EC25) and $\tilde{\pi}_\tau(v)$ is continuous in v, there exists $\delta^+ > \delta$ so that $\tilde{\pi}_\tau(v) \leq 0$.

Consider therefore a $\delta, \delta^+ > 0$ such that $\tilde{\pi}_s(\hat{v}) < \tilde{\pi}_s(v), \forall s \notin \{t, \tau\}$, and $\tilde{\pi}_s(\hat{v}) \leq 0, \forall s \in \{t, \tau\}$. Because v is feasible, $\hat{v}$ is feasible in problem (RS-PFP). Moreover, the objective under $(T, \hat{v})$ is:

$$\begin{aligned}\Pi(T, \hat{v}) &= \sum_{f \in \mathcal{F}} \hat{v}_f - \sum_{s \in \mathcal{T}} \max(\tilde{\pi}_s(\hat{v}), 0) - \sum_{t \in T} \sigma_t \geq \sum_{f \in \mathcal{F}} v_f + n_h(\delta^+ - \delta) - \sum_{s \in \mathcal{T}} \max(\hat{\pi}_s(v), 0) - \sum_{t \in T} \sigma_t \\ &= \Pi(T, v) + (n_h - \min(\epsilon_h, K - n_h))\delta > \Pi(T, v), \end{aligned}$$

where the inequality contradicts the optimality of (T, v).

Sub-case (e-2): $\tilde{\pi}_t(v) > 0$. Consider a new vector of margins $\hat{v}$ given by:

$$\hat{v}_f = \begin{cases} v_f - \delta & \text{if } f \in \mathcal{H}_t \\ v_f + \delta \frac{\epsilon_s}{n_s - \kappa_s} & \text{if } f \in \mathcal{H}_s \text{ and } s \neq t, \end{cases} \tag{EC26}$$

for some $\delta > 0$ with $\delta \leq \frac{\bar{v} - \bar{v}_2}{1 + \epsilon_s / \min(n_s, K - \epsilon_s)}$ for all $s \in \mathcal{T}$ s.t $s \neq t$. Notice that if $s \neq t$ and $f \in \mathcal{H}_s$, we have that:

$$\delta\left(1 + \frac{\epsilon_s}{\min(n_s, K - \epsilon_s)}\right) \leq \bar{v} - \bar{v}_2 \leq \bar{v} - v_f \quad \Rightarrow \quad \bar{v} - \delta \geq v_f + \delta \frac{\epsilon_s}{\min(n_s, K - \epsilon_s)},$$

which implies that $\hat{v}^{\max} = \bar{v} - \delta$ and $\hat{v}_2^{\max} = v^\tau + \delta \frac{\epsilon_\tau}{\min(n_\tau, K-\epsilon_\tau)}$ for some $\tau \neq t$. From (EC24), we have:

$$\begin{aligned}
\tilde{\pi}_s(\hat{v}) &= \min(n_s, K-\epsilon_s)\left(v^s + \delta\frac{\epsilon_s}{\min(n_s, K-\epsilon_s)}\right) + \epsilon_s(\bar{v}-\delta) - \sigma_s = \tilde{\pi}_s(v), \quad \forall\, s \neq t \\
\tilde{\pi}_t(\hat{v}) &= n_h(\bar{v}-\delta) + \left(v^\tau + \delta\frac{\epsilon_\tau}{\min(n_\tau, K-\epsilon_\tau)}\right)\min(\epsilon_h, K-n_h) \\
&\leq n_h(\bar{v}-\delta) + \left(\bar{v}_2 + \delta\frac{\epsilon_\tau}{\min(n_\tau, K-\epsilon_\tau)}\right)\min(\epsilon_h, K-n_h) = \tilde{\pi}_t(v) - \delta n_h + \delta\min(\epsilon_h, K-n_h)\frac{\epsilon_\tau}{\min(n_\tau, K-\epsilon_\tau)}.
\end{aligned}$$

This shows that $(T, \hat{v})$ remains feasible in problem (RS-PFP). Moreover, the difference in objective values between $(T, \hat{v})$ and (T, v) satisfies:

$$\begin{aligned}
\Pi(T, \hat{v}) - \Pi(T, v) &= \left(\sum_{f\in\mathcal{F}} \hat{v}_f - \sum_{s\in\mathcal{T}} \max(\tilde{\pi}_s(\hat{v}), 0) - \sum_{s\in T}\sigma_s\right) - \left(\sum_{f\in\mathcal{F}} v_f - \sum_{s\in\mathcal{T}} \max(\tilde{\pi}_s(v), 0) - \sum_{s\in T}\sigma_s\right) \\
&= \sum_{s\in\mathcal{T}: s\neq t} \delta n_s \frac{\epsilon_s}{\min(n_s, K-\epsilon_s)} - \delta\min(\epsilon_h, K-n_h)\frac{\epsilon_\tau}{\min(n_\tau, K-\epsilon_\tau)}.
\end{aligned}$$

To prove that the difference above is positive, consider first the case $\tau \in \mathcal{T}_\ell$, when we have:

$$\sum_{s\in\mathcal{T}: s\neq t} \delta n_s \frac{\epsilon_s}{\min(n_s, K-\epsilon_s)} \geq \delta|\mathcal{H}[\mathcal{T}_\ell]|\frac{\epsilon_\ell}{n_\ell} > \delta\epsilon_h\frac{\epsilon_\ell}{n_\ell} \geq \delta\min(\epsilon_h, K-n_h)\frac{\epsilon_\tau}{\min(n_\tau, K-\epsilon_\tau)},$$

where we use the fact that $\min(n_\tau, K-\epsilon_\tau) = \min(n_\ell, K-\epsilon_\ell) = n_\ell$. If $\tau \in \mathcal{T}_h$, the inequality follows from:

$$\sum_{s\in\mathcal{T}: s\neq t} \delta n_s \frac{\epsilon_s}{\min(n_s, K-\epsilon_s)} \geq \delta n_\tau \frac{\epsilon_\tau}{\min(n_\tau, K-\epsilon_\tau)} = \delta n_h \frac{\epsilon_\tau}{\min(n_\tau, K-\epsilon_\tau)} > \delta\min(\epsilon_h, K-n_h)\frac{\epsilon_\tau}{\min(n_\tau, K-\epsilon_\tau)},$$

where the last inequality follows because $n_h \geq \epsilon_h$ from Assumption 4-(b).

Part (iv). We first prove "$\subseteq$" in (EC22). Take (T, v) optimal in (RS-PFP) with $v \in \mathcal{S}^{\text{mg-eq}}$. (EC22a) restates that $v \in \mathcal{S}^{\text{mg-eq}}$ and defines $v^t = v_f$ for $f \in \mathcal{H}_t$. We prove that (T, v) verifies (EC22b). First, we rewrite (RS-PFP) by maximizing over (T, v) optimal with $v \in \mathcal{S}^{\text{mg-eq}}$; because for such (T, v), (EC20) holds by part (iii), (RS-PFP) has the same optimal value as:

$$\max_{T\subseteq\mathcal{T}, \{v^t\in\mathbb{R}\}_{t\in\mathcal{T}}, \bar{v}\in\mathbb{R}} \ \sum_{t\in\mathcal{T}} v^t n_t - \sum_{t\in\mathcal{T}} \max\left(\min(n_t, K-\epsilon_t)v^t + \epsilon_t\bar{v} - \sigma_t, 0\right) - \sum_{t\in T}\sigma_t \tag{EC27a}$$

$$\text{s.t. } 0 \geq \min(n_t, K-\epsilon_t)v^t + \epsilon_t\bar{v} - \sigma_t, \quad \forall t \in T^c \cap \mathcal{T}_h \tag{EC27b}$$

$$|T| \geq |\mathcal{F}|/K, \quad \bar{v} = \max_{t\in\mathcal{T}} v^t. \tag{EC27c}$$

Because (T, v) is optimal in (RS-PFP), $\{v^t\}_{t\in\mathcal{T}}$ and $\bar{v}$ must be optimal in (EC27). Note that for fixed $\bar{v}, T$, the problem of choosing v^t above is separable over $t \in \mathcal{T}$. We first show that the optimal values for v^t as a function of $T, \bar{v}$ are:

$$v^t \in \left[\min\left(\bar{v}, (\sigma_\ell - \epsilon_\ell\bar{v})/n_\ell\right), \ \bar{v}\right] \qquad \text{if } t \in \mathcal{T}_\ell \tag{EC28a}$$

$$v^t \in \left[\min\left(\bar{v}, (\sigma_h - \epsilon_h\bar{v})/n_h\right), \ \bar{v}\right] \qquad \text{if } t \in T \cap \mathcal{T}_h \text{ and } n_h + \epsilon_h \leq K \tag{EC28b}$$

$$v^t = \bar{v} \qquad \text{if } t \in T \cap \mathcal{T}_h \text{ and } n_h + \epsilon_h > K \tag{EC28c}$$

$$v^t = \min\left(\bar{v}, \ \left(\sigma_h - \epsilon_h\bar{v}\right)/\min(n_h, K-\epsilon_h)\right) \qquad \text{if } t \in T^c \cap \mathcal{T}_h \tag{EC28d}$$

Consider the following cases:

(a) $t \in \mathcal{T}_\ell$. Then, $\min(n_\ell, K-\epsilon_\ell) = n_\ell$ and the objective (EC27a) is strictly increasing in v^t if $n_t v^t + \epsilon_t\bar{v} - \sigma_t < 0 \Leftrightarrow v^t \leq \frac{\sigma_\ell - \epsilon_\ell\bar{v}}{n_\ell}$, and is independent of v^t otherwise. Because $v^t \leq \bar{v}$ must hold, any v^t satisfying (EC28a) is optimal.

(b) $t \in T \cap \mathcal{T}_h$ and $n_h + \epsilon_h \le K$. Then, $\min(n_t, K - \epsilon_t) = n_t$ and the analysis is analogous to (a).

(c) $t \in T \cap \mathcal{T}_h$ and $n_h + \epsilon_h > K$. Then, $\min(n_t, K - \epsilon_t) = K - \epsilon_t < n_t$, so the objective is strictly increasing in v^t on $[0, \bar{v}]$ and it is optimal to set $v^t = \bar{v}$, as in (EC28c).

(d) $t \in T^c \cap \mathcal{T}_h$. The objective is strictly increasing in v^t on $[0, \bar{v}]$ by the same argument as in (c). So it is optimal to increase v^t until either (EC27b) becomes binding or $v^t = \bar{v}$, which shows that (EC28d) is optimal.

We show that (EC28a)-(EC28d) are equivalent to (EC22c)-(EC22f) after analyzing optimal choices for $\bar{v}$. We first find a closed form expression for the objective (EC27a). If we use $v^t = \bar{v}$ for $t \in \mathcal{T}_\ell \cup (T \cap \mathcal{T}_h)$ (which is optimal in (EC28a), (EC28b) and (EC28c)) and the optimal value (EC28d) for $t \in T^c \cap \mathcal{T}_h$, we can rewrite (EC27a) as:

$$\sum_{t \in \mathcal{T}_\ell \cup (T \cap \mathcal{T}_h)} n_t \bar{v} + \sum_{t \in T^c \cap \mathcal{T}_h} n_h \min\left(\bar{v}, \frac{\sigma_h - \epsilon_h \bar{v}}{\min(n_h, K - \epsilon_h)}\right) - \sum_{t \in \mathcal{T}_\ell \cup (T \cap \mathcal{T}_h)} \left((\min(n_t + \epsilon_t, K))\bar{v} - \sigma_t\right)^+ - \sum_{t \in T} \sigma_t. \quad \text{(EC29)}$$

Note that this expression already uses constraint (EC27b) to simplify the third sum in the expression above to only contain terms $t \in \mathcal{T}_\ell \cup (T \cap \mathcal{T}_h)$. To simplify this further, consider a typical term in the second sum:

$$\begin{aligned} n_h \min\left(\bar{v}, \frac{\sigma_h - \epsilon_h \bar{v}}{\min(n_h, K - \epsilon_h)}\right) &= n_h \bar{v} + n_h \min\left(0, \frac{\sigma_h - \epsilon_h \bar{v}}{\min(n_h, K - \epsilon_h)} - \bar{v}\right) \\ &= n_h \bar{v} - \max(\min(n_t + \epsilon_t, K)\bar{v} - \sigma_h, 0) - \frac{\max(0, n_h + \epsilon_h - K)}{\min(n_h, K - \epsilon_h)} \max\big(0, \min(n_h + \epsilon_h, K)\bar{v} - \sigma_h\big). \end{aligned}$$

where we skip some algebra for brevity. Replacing this into (EC27a) and using $\sum_{t \in \mathcal{T}} n_t = |\mathcal{F}|$ then yields exactly:

$$\begin{aligned} &|\mathcal{F}|\bar{v} - \sum_{t \in \mathcal{T}} \big(\min(n_t + \epsilon_t, K)\bar{v} - \sigma_t\big)^+ - \frac{(n_h + \epsilon_h - K)^+}{\min(n_h, K - \epsilon_h)} \sum_{t \in T^c \cap \mathcal{T}_h} \big(\min(n_h + \epsilon_h, K)\bar{v} - \sigma_h\big)^+ - \sum_{t \in T} \sigma_t \\ &\quad = |\mathcal{F}|\, v_{\max} - |\mathcal{T}_\ell|\big[A_\ell v_{\max} - \sigma_\ell\big]^+ - |\mathcal{T}_h|\big[A_h v_{\max} - \sigma_h\big]^+ - \big(|\mathcal{T}_h| - x_h\big)\delta_h\big[A_h v_{\max} - \sigma_h\big]^+ - x_\ell \sigma_\ell - x_h \sigma_h, \quad \text{(EC30)} \end{aligned}$$

where $A_\ell = n_\ell + \epsilon_\ell$, $A_h = \min(n_h + \epsilon_h, K)$, $x_\theta = |T \cap \mathcal{T}_\theta|, \forall \theta \in \{\ell, h\}$, $\delta_h = \frac{\max(0, n_h + \epsilon_h - K)}{\min(n_h, K - \epsilon_h)}$. The choice of T translates into the number of intermediaries $x_\theta = |T \cap \mathcal{T}_\theta|$ of each type to match. Any x_ℓ, x_h satisfying the following constraints is feasible:

$$0 \le x_\ell \le |\mathcal{T}_\ell|, \qquad 0 \le x_h \le |\mathcal{T}_h|, \qquad x_\ell + x_h \ge |\mathcal{F}|/K, \qquad x_\ell, x_h \text{ integer.}$$

Assumption 4(e) and the first three feasibility constraints imply that $x_\ell + x_h \ge |\mathcal{F}|/K > |\mathcal{T}_h| \ge x_h$, so in any feasible solution we must have $x_\ell > 0$. Moreover, because (EC30) is strictly decreasing in x_ℓ, at optimality we must have $x_\ell + x_h = |\mathcal{F}|/K$. This together with $|\mathcal{T}_\ell| > |\mathcal{F}|/K$ (which holds by Assumption 4(e)) implies that $x_\ell^\star \in (0, |\mathcal{T}_\ell|)$ and optimal (basic feasible) solutions exist maximizing (EC30) where either $(x_\ell^\star, x_h^\star) = (|\mathcal{F}|/K, 0)$ or $(x_\ell^\star, x_h^\star) = \big(|\mathcal{F}|/K - |\mathcal{T}_h|, |\mathcal{T}_h|\big)$. Therefore, relaxing the integrality requirement on x_ℓ, x_h does not affect the optimal solution. Replacing $x_\ell + x_h = |\mathcal{F}|/K$ into (EC30) then yields $\Pi(x_h, \bar{v})$ in (EC21), proving that $(x_h, \bar{v})$ must be optimal in problem (EC21).

We determine the optimal $\bar{v}$. $\Pi(x_h, \bar{v})$ is piece-wise linear and concave in $\bar{v}$, and its subgradient $\frac{\partial \Pi(x_h, \bar{v})}{\partial \bar{v}}$ is discontinuous at $e_\theta = \sigma_\theta / A_\theta, \theta \in \{h, \ell\}$. From Assumption 4(d), these points satisfy:

$$e_h = \frac{\sigma_h}{A_h} = \frac{\sigma_h}{\min(n_h + \epsilon_h, K)} < e_\ell = \frac{\sigma_\ell}{A_\ell} = \frac{\sigma_\ell}{n_\ell + \epsilon_\ell}. \quad \text{(EC31)}$$

For $\bar{v} < \sigma_h / A_h$, we have $\frac{\partial \Pi(x_h, \bar{v})}{\partial \bar{v}} = |\mathcal{F}| > 0$. For $\bar{v} > \sigma_\ell / A_\ell$, we have:

$$\frac{\partial \Pi(x_h, \bar{v})}{\partial \bar{v}} \le |\mathcal{F}| - |\mathcal{T}_h| A_h - |\mathcal{T}_\ell| A_\ell < |\mathcal{F}| - |\mathcal{T}_h| n_h - |\mathcal{T}_\ell| n_\ell = 0,$$

where the second inequality follows because $A_h \geq n_h$ because $K \geq n_h$. Therefore, the maximum occurs either at $\bar{v} \in \{e_h, e_\ell\}$ or at every value in $\bar{v} \in [e_h, e_\ell]$ (in a degenerate case). Moreover, note that:

$$\bar{v} \geq \frac{\sigma_h - \epsilon_h \bar{v}}{\min(n_h, K - \epsilon_h)} \iff \bar{v} \geq e_h \quad \text{and} \quad \bar{v} \leq \frac{\sigma_\ell - \epsilon_\ell \bar{v}}{n_\ell} \iff \bar{v} \leq e_\ell, \tag{EC32}$$

which always hold because $\bar{v} \in [e_\ell, e_h]$ at optimality. Therefore, conditions (EC28a)-(EC28d) are equivalent to (EC22c)-(EC22f).

This proves the inclusion $\subseteq$ in (EC22). To prove inclusion $\supseteq$, note that for any v that satisfies (EC22c)-(EC22f), the optimal profit $\tilde{\pi}_t(v)$ in the deviation problem (17) is exactly given by (EC20), because there are at least $|\mathcal{T}_\ell|$ farmers with margin $\bar{v}$, and Assumption 4-(g) implies $|\mathcal{T}_\ell| \geq K$, so $\mathcal{H}_t^c$ always contains at least ϵ_t farmers with margin $\bar{v}$. Because (EC20) holds, we can repeat the calculations above to show that any (T, v) that satisfies (EC22a)-(EC22f) is optimal for (RS-PFP).

Part (v). Take any optimal solution $(T, v) \in \mathcal{S}^{\star,\text{mg-eq}}$ for (RS-PFP). We know that $v_f \leq e_\ell \leq p - c_f$, where the second inequality follows from Assumption 4(f). This implies that the constraint (18d)–which was relaxed in $(\text{S-PFP})_2$ to obtain (RS-PFP)–holds for (T, v). Moreover, we claim that $\tilde{\pi}_t(v) \leq 0, \forall t \in T^c$. This holds for $t \in T^c \cap \mathcal{T}_h$ because (T, v) is feasible in (RS-PFP). For $t \in \mathcal{T}_\ell$, note that the expression of profit $\tilde{\pi}_t(v)$ from (EC20) satisfies:

$$\tilde{\pi}_t(v) = \min(n_t, K - \epsilon_t)\bar{v} + \epsilon_t \bar{v} - \sigma_t = (n_\ell + \epsilon_\ell)\bar{v} - \sigma_\ell \leq (n_\ell + \epsilon_\ell)e_\ell - \sigma_\ell = 0, \ \forall t \in \mathcal{T}_\ell, \tag{EC33}$$

where the inequality follows because $\bar{v} \leq e_\ell$. This implies that (T, v) is feasible for $(\text{S-PFP})_2$. Because (RS-PFP) was a relaxation of $(\text{S-PFP})_2$, both problems admit (T, v) as an optimal solution, implying the optimal values are also equal.

Part (vi). With $\mathcal{A} \in \{\mathcal{F}, \mathcal{T}\}$, note that an extremal (maximum/minimum) value of $\mathcal{W}^{\mathcal{A}}$ over all optimal solutions (T, v) to $(\text{S-PFP})_2$ is achieved with $(T, v) \in \mathcal{S}^{\star,\text{mg-eq}}$. This follows because (RS-PFP) is a relaxation of $(\text{S-PFP})_2$ and extremal values of $\mathcal{W}^{\mathcal{A}}$ over optimal solutions (T, v) to (RS-PFP) are achieved with $v \in \mathcal{S}^{\text{mg-eq}}$ by part (ii), and (RS-PFP) and $(\text{S-PFP})_2$ admit identical optimal solutions that are also margin-equalizing, by part (v).

Therefore, we characterize extremal welfare $\mathcal{W}^{\mathcal{A}}$ over solutions $(T, v) \in \mathcal{S}^{\star,\text{mg-eq}}$. Consider maximizing $\mathcal{W}^{\mathcal{F}} = \mathbf{1}^\top(p - c - v)$, which is equivalent to minimizing $\mathbf{1}^\top v$. Because $v_f = v^t$ for all $f \in \mathcal{H}_t$ and all $t \in \mathcal{T}$, the optimal choice of v involves setting each v^t to the smallest possible value in (EC22c)-(EC22f), for any $t \in \mathcal{T}$, which leads us to the objective appearing in the maximization in (EC23a). Maximizing this expression over choices $(x_h, \bar{v}) \in X^\star$ then yields $\overline{\mathcal{W}^{\mathcal{F}}}$ in (EC23a). The expression for the minimum farmer welfare $\underline{\mathcal{W}^{\mathcal{F}}}$ in (EC23b) involves maximizing $\mathbf{1}^\top v$ and follows from similar ideas.

To calculate extremal intermediary welfare $\mathcal{W}^{\mathcal{T}}$, recall that $\tilde{\pi}_t(v) \leq 0, \forall t \in \mathcal{T}_\ell \cup (T^c \cap \mathcal{T}_h)$ for all $(T, v) \in \mathcal{S}^{\star,\text{mg-eq}}$, which follows from (EC33) and from the feasibility of (T, v) for $(\text{S-PFP})_2$. Therefore, because $\pi_t^\star = \max(\tilde{\pi}_t(v), 0)$, only intermediaries $t \in T \cap \mathcal{T}_h$ make positive profit and the intermediary welfare is $\sum_{t \in T \cap \mathcal{T}_h} \max(\tilde{\pi}_t(v), 0)$. Because $\tilde{\pi}_t(v)$ is increasing in v^t as shown by (EC20), in extremizing $\sum_{t \in T \cap \mathcal{T}_h} \pi_t(v)$, it suffices to consider the maximum and minimum margins v^t from (EC22d) and (EC22e). Note that for $v^t = \bar{v}$ we have:

$$\tilde{\pi}_t(v) = \min(n_h, K - \epsilon_h)\bar{v} + \epsilon_h \bar{v} - \sigma_h = \min(n_h + \epsilon_h, K)\bar{v} - \sigma_h.$$

Similarly, for $v^t = \frac{\sigma_h - \epsilon_h \bar{v}}{\min(n_h, K - \epsilon_h)}$ we have

$$\tilde{\pi}_t(v) = \min(n_h, K - \epsilon_h)\frac{\sigma_h - \epsilon_h \bar{v}}{\min(n_h, K - \epsilon_h)} + \epsilon_h \bar{v} - \sigma_h = 0.$$

Replacing these values yields equations for the maximum an minimum welfare (EC23c) and (EC23d), respectively.

Lastly, we show that the extremal welfare levels in (EC23a)-(EC23d) are achieved for $x_h \in \{0, |\mathcal{T}_h|\}$ and $\bar{v} in \{e_h, e_\ell\}$. We first characterize the set $X^\star$ of optimal solutions to (EC21). Part (iv) already argued that for a fixed x_h, the maximizer of (EC21) occurs either at $\bar{v} \in \{e_h, e_\ell\}$ or for *every* $\bar{v} \in [e_h, e_\ell]$ (in a degenerate case); and similarly, that for any $\bar{v}$, it is optimal to choose either $x_h \in \{0, |\mathcal{T}_h|\}$ or the whole interval $x_h \in [0, |\mathcal{T}_h|]$ (in a degenerate case). This shows that the set X^* is a face of the polytope $[0, |\mathcal{T}_h|] \times [e_h, e_\ell]$ (it is either equal the full square $[0, |\mathcal{T}_h|] \times [e_h, e_\ell]$ or one sides of that square or a square of that square). The, noting that all the objectives that appear in (EC23a)-(EC23d) are bilinear functions of $x_h, \bar{v}$, there always exists an optimal value in (EC23a)-(EC23d) that belongs to the extreme points of $X^\star$. And because $X^\star$ is a face of the polytope $[0, |\mathcal{T}_h|] \times [e_h, e_\ell]$, an optimum exists at an extreme point of the polytope.

□

C.2. Auxiliary Results

DEFINITION 2. The function $\text{avg} : \mathbb{R}^n \times 2^n \to \mathbb{R}^n$ is defined as:

$$\text{avg}(x, S) = \begin{cases} \frac{1}{|S|} \sum_{j \in S} x_j & \text{if } i \in S \\ x_i & \text{if } i \notin S \end{cases}, \ \forall x \in \mathbb{R}^n, \ \forall S \subseteq \{1, \ldots, n\}. \tag{EC34}$$

Given vector x and a subset of coordinates S, $\text{avg}(x, S)$ returns a vector that is identical to x except for the components $\{x_i : i \in S\}$, whose value is averaged.

PROPOSITION 6. *Consider a convex function $f : \mathbb{R}^n \to \mathbb{R}$ that is convex and symmetric in $\{x_i : i \in S\}$ for some $S \subseteq \{1, \ldots, n\}$, that is, $f(x) = f(x^\sigma)$ for any $x^\sigma \in \mathbb{R}^n$ satisfying $x^\sigma_i = x_i$ for $i \notin S$ and $x^\sigma_i = x_{\sigma(i)}$ for $i \in S$ for some permutation σ of S. Then, $f(x) \geq f(\text{avg}(x, S))$.*

Proof of Proposition 6. Fix any value for x_i with $i \notin S$ and note that $\text{avg}(x, S)$ also equals the arithmetic average of all vectors obtained by permuting the coordinates in S, that is $\text{avg}(x, S) = \frac{1}{|\Sigma|} \sum_{\sigma \in \Sigma} x^\sigma$, where Σ is the set of all permutations of $\{1, \ldots, n\}$ and $x^\sigma \in \mathbb{R}^n$ is given by $x^\sigma_i = x_i$ for $i \notin S$ and $x^\sigma_i = x_{\sigma(i)}$ for $i \in S$, for any permutation $\sigma \in \Sigma$. Then, we readily have:

$$f(\bar{x}^S) = f\left(\frac{1}{|\Sigma|} \sum_{\sigma \in \Sigma} x^\sigma\right) \leq \frac{1}{|\Sigma|} \sum_{\sigma \in \Sigma} f(x^\sigma) = f(x),$$

where the inequality follows by convexity and the last equality follows by the symmetry assumption. □

We next leverage Proposition 6 to prove structural properties on the function $\tilde{\pi}_t(v)$ defined in (17).

PROPOSITION 7. *Consider a set of intermediaries $S \subseteq \mathcal{T}$ of the same type ($S \subseteq \mathcal{T}_\theta$ for some $\theta \in \{\ell, h\}$) and let $\mathcal{H}[S] := \cup_{t \in S} \mathcal{H}_t$ denote all they serve collectively in the status quo. With $\tilde{\pi}_t(v)$ defined in* (17)*, we have:*

$$\max(\tilde{\pi}_t(v), 0) \geq \max\Big(\tilde{\pi}_t\big(\text{avg}(v, \mathcal{H}[S])\big), 0\Big), \quad \forall t \in \mathcal{T} \setminus S \tag{EC35a}$$

$$\sum_{t \in S} \max(\tilde{\pi}_t(v), 0) \geq \sum_{t \in S} \max\Big(\tilde{\pi}_t\big(\text{avg}(v, \mathcal{H}[S])\big), 0\Big). \tag{EC35b}$$

where avg(,) *is defined in Definition 2.*

Proof of Proposition 7. To facilitate following the proof, we repeat the definition $\tilde{\pi}_t(v)$ from (17):

$$\tilde{\pi}_t(v) := -\sigma_t + \sup_{\mu_t \in [0,1]^{\mathcal{F}}} \left\{ \mu_t^\top v \ : \ \sum_{f \in \mathcal{H}_t^c} \mu_{tf} \leq \epsilon_t, \ \sum_{f \in \mathcal{F}} \mu_{tf} \leq K \right\}, \ \text{where} \ \ v := p - r - c.$$

We first show (EC35a). First, note from the definition of $\tilde{\pi}_t(v)$ that $\tilde{\pi}_t$ is convex in v and is symmetric in variables $\{v_f : f \in \mathcal{H}_t^c\}$, and the function $\max(\tilde{\pi}_t(\cdot), 0)$ has the same properties. Because $\mathcal{H}_t \cap \mathcal{H}_{t'} = \emptyset$ for any $t \neq t' \in \mathcal{T}$, we have that $\mathcal{H}[S] \subseteq \mathcal{H}_t^c$ for any $t \in \mathcal{T} \setminus S$, so an application of Proposition 6 to the function $\max(\tilde{\pi}_t(\cdot), 0)$ with the set of coordinates $\mathcal{H}[S]$ yields the result.

To prove (EC35b), define a new vector v' by averaging the components $\{v_f : f \in \mathcal{H}_t\}$ for every $t \in S$:

$$v'_f = \begin{cases} \frac{1}{|\mathcal{H}_t|} \sum_{g \in \mathcal{H}_t} v_g & \text{if } f \in \mathcal{H}_t \text{ for some } t \in S \\ v_f, & \text{otherwise} \end{cases}$$

Fix $t \in S$. Note that the function $\max\big(0, \tilde{\pi}_t(v)\big)$ is convex in v and for every $s \in S$, it is symmetric in the components $\{v_f : f \in \mathcal{H}_s\}$. Therefore, by applying Proposition 6 to this function iteratively, for every $s \in S$, with a choice of coordinates $\mathcal{H}_s$, we can conclude that

$$\max(\tilde{\pi}_t(v'), 0) \leq \max(\tilde{\pi}_t(v), 0), \ \forall t \in S.$$

Next, define a lifting function $L : \mathbb{R}^S \to \mathbb{R}^{\mathcal{F}}$ as $[L(x)]_f = x_s$ for every $f \in \mathcal{H}_s$ and $s \in S$, and $[L(x)]_f = v_f$ for $f \in \mathcal{H}[S]^c$. Intuitively, $L(\cdot)$ takes as input a value x_s for each margin for $s \in S$ and returns a vector of margins for all farmers shaped like v', where the margin of every farmer $f \in H_s$ is given by x_s (for every $s \in S$) and all other margins are set to v_f. Now, consider the function $g : \mathbb{R}^S \to \mathbb{R}$ defined as:

$$g(x) := \sum_{t \in S} \max\big(\tilde{\pi}_t(L(x)), 0\big).$$

It can be readily checked that g is convex in x. Moreover, because all intermediaries $t \in S$ have the same type, the function $g(x)$ is also symmetric in $\{x_s : s \in S\}$. Therefore, Proposition 6 yields $g(x) \geq g\big(\text{avg}(x, S)\big)$. To conclude the proof, note that $\text{avg}(v, \mathcal{H}[S]) = L(\text{avg}(x, S))$ (because $|\mathcal{H}_s|$ takes the same value for all $s \in S$), so we have:

$$\sum_{t \in S} \max(\tilde{\pi}_t(v), 0) \geq \sum_{t \in S} \max(\tilde{\pi}_t(v'), 0) = g(x) \geq g(\text{avg}(x, S)) = \sum_{t \in S} \max\Big(\tilde{\pi}_t\big(\text{avg}(v, \mathcal{H}[S])\big), 0\Big). \quad \square$$

Appendix D: Extensions

D.1. Fractional Collections

Our model can be extended to allow fractional fruit collection, motivated either by the platform's desire to improve logistical efficiency or by a more conservative formulation that allows intermediaries to deviate with partial amounts of fruit. For platform cost efficiency motivations, fractional collections allow multiple intermediaries share a farmer's harvest. In this case, the Matching Oracle from Assumption 3 can be adapted either by (i) modeling Q_f units of fruit as Q_f identical farmers, each supplying one unit, or (ii) adopting a continuous formulation where the oracle directly solves over fractional quantities. Case (ii), under the setting of transportation costs coming from routing costs, reduces to the well-studied Split Delivery VRP (Archetti and Speranza 2008).

When extending the model for deviation motivations, intermediaries may collect only part of a farmer's harvest when deviating. Let $d_t \in \mathbb{R}^{\mathcal{F}}$ denote the (potentially franctional) quantities collected by intermediary t when deviating, and let $u_t \in \{0, 1\}^{\mathcal{F}}$ denote the candidate vector that indicates the farmers that agree to potentially deviate with intermediary t. We constrain $d_{tf} = 0$ whenever $u_{tf} = 0$ and the deviation profit, under fractional pickups, becomes

$$\hat{\pi}_t(r) := \sup_{\mathbb{P}_t \in \mathcal{P}_t} \mathbb{E}_{u_t \sim \mathbb{P}_t} \left[\max_{d_t \in \mathcal{I}_t : d_t \leq u_t} \big((pq - r)^\top d_t - c_t(d_t) \big) \right], \tag{EC36}$$

where $\mathcal{I}_t$ is the feasible set and $c_t(d_t)$ the cost of collecting d_t. To preserve compatibility with our algorithms, $\mathcal{I}_t$ must remain discrete, which can be ensured by treating each unit of a farmer's harvest as distinct.

D.2. Quantity uncertainty

We extend the platform problem to account for uncertainty in the quantities of palm fruit produced by farmers. Let the vector of production quantities be denoted by q, drawn from a known distribution $\mathbb{Q}$ observed by both the farmers and the platform. The realized quantity becomes known when the intermediary arrives at the farmer's plantation. We consider a capacitated setting in which each intermediary can transport at most Q units of fruit.

To incorporate this uncertainty, we redefine the set of feasible schedules for intermediary t as

$$\mathcal{I}_t = \left\{ s_t \in \{0,1\}^{\mathcal{F}} : \rho(Q^\top s_t) \leq \beta \right\},$$

where $\rho(\cdot)$ is a risk measure and β is an acceptable risk level. For instance, setting $\beta = K$ and letting ρ be the Value-at-Risk at level α enforces that the probability of exceeding capacity K is at most α. Alternatively, ρ could represent the Expected Shortfall at level α, or the maximum value in the support of the distribution.

The tractability of Oracle (5) and Oracle (12) in this setting depends on the structure of $\mathcal{I}_t$. For instance, when ρ is the Value-at-Risk at level α, the convex hull of $\mathcal{I}_t$ is tractable in many settings. Similarly, when the realizations of Q are known to lie within a polyhedral uncertainty set, the problem falls within the scope of some work in the robust vehicle routing literature (see, e.g., Gounaris et al. 2013).

D.3. Additional Ambiguity Sets

We discuss two additional examples of classes of ambiguity sets compatible with our framework.

Example 5 (Data-Driven ϕ-divergence Ambiguity). For $\mathbb{P}_1, \mathbb{P}_2 \in \Delta_{\mathcal{F}}$, define the ϕ-divergence $\mathbb{I}_\phi$ as

$$\mathbb{I}_\phi(\mathbb{P}_1, \mathbb{P}_2) := \sum_{u \in \{0,1\}^{\mathcal{F}}} \mathbb{P}_2(u) \cdot \phi\left(\frac{\mathbb{P}_1(u)}{\mathbb{P}_2(u)}\right), \tag{EC37}$$

where $\phi(x)$ is convex for $x \geq 0$ and satisfies $\phi(1) = 0$, and we define $0 \cdot \phi(a/0) := a \cdot \lim_{x\to\infty}(\phi(x)/x)$ for any $a > 0$, and $0 \cdot \phi(0/0) := 0$. Following Ben-Tal et al. (2013), we construct our ambiguity sets as:

$$\mathcal{P}_t^\phi = \left\{ \mathbb{P}_t \in \Delta_{\mathcal{F}} : \mathbb{I}_\phi(\mathbb{P}_t, \hat{\mathbb{P}}_t) \leq \epsilon_t \right\}. \tag{EC38}$$

Selecting different ϕ-divergences allows encoding qualitatively different ambiguity sets and encoding the intermediary's risk attitude. For example, Burg entropy, $\phi(x) = -\log x + x - 1$, produces all distributions whose likelihood ratio relative to the nominal law $\hat{\mathbb{P}}_t$ is bounded by a monotone function of the radius ϵ_t, thereby controlling statistical proximity. More generally, replacing the risk-neutral expectation in the deviation profit with a coherent risk functional is achieved by an appropriate choice of ϕ. For instance, choosing $\phi(x) = 0$ for $0 \leq x \leq 1/(1-\beta)$ and $\phi(x) = \infty$ otherwise makes the resulting expression coincide with the Conditional Value at Risk at level β, and every coherent risk measure admits such a representation (see Ben-Tal et al. (2013) and Bayraksan and Love (2015) for a detailed treatment).

The next result shows that ambiguity sets based on ϕ-divergence satisfy Assumption 2.

Proposition 8 (**ϕ-Divergence**). *If $\mathcal{P}_t = \mathcal{P}_t^\phi$ defined in* (EC38)*, the optimal profit $\hat{\pi}_t(r)$ in* (1) *can be obtained by solving the following convex optimization problem in variables μ, η and $\{\kappa^u\}_{u \in \mathrm{supp}(\hat{\mathbb{P}}_t)}$:*

$$\hat{\pi}_t(r) = \inf_{\mu,\eta \geq 0, \kappa^u} \left(\mu + \eta \cdot \epsilon_t + \eta \sum_{u \in \{0,1\}^{\mathcal{F}} : \hat{\mathbb{P}}_t(u) > 0} \hat{\mathbb{P}}_t(u) \cdot \phi^\star\left(\frac{\kappa^u}{\eta}\right) \right) \tag{EC39a}$$

$$\text{s.t. } (p \cdot q - r)^\top d_t - c_t(d_t) - \mu \leq \eta \cdot \bar{s} \quad \forall d_t \in \mathcal{I}_t \tag{EC39b}$$

$$(p \cdot q - r)^\top d_t - c_t(d_t) - \mu \leq \kappa^u \quad \forall d_t \in \mathcal{I}_t : d \leq u,\ \forall u \in \{0,1\}^{\mathcal{F}} : \hat{\mathbb{P}}_t(u) > 0, \tag{EC39c}$$

where $\bar{s} := \lim_{x\to\infty}(\phi(x)/x)$ *and* $\phi^\star$ *is the convex conjugate of* ϕ*,* $\phi^\star(s) = \sup_{t\geq 0}(st - \phi(t))$*. Moreover, constraint* (EC39c) *can be separated with* $|\operatorname{supp}(\hat{\mathbb{P}}_t)|$ *calls to oracle* (5)*.*

The proof follows directly from Corollary 1 in Ben-Tal et al. (2013) and is omitted. Depending on the choice of ϕ, problem (EC39) can be a linear, second-order conic, or general convex optimization problem.

Example 6 (Known Marginals). Because collecting historical data on transactions between every intermediary and every farmer may be challenging, one could instead calibrate a Machine Learning (ML) model to predict the probability μ_{tf} that t and f are willing to transact together, based on features from intermediary t and farmer f. Moreover, bootstrapping or Bayesian methods could be used to construct a convex confidence interval $P_t(\epsilon_t) \subseteq \mathbb{R}^{\mathcal{F}}$ that contains the vector of real probabilities μ_t with confidence ϵ_t. Then, we can consider the following ambiguity set:

$$\mathcal{P}_t^{\text{corr}} = \left\{\mathbb{P}_t \in \Delta_{\mathcal{F}} :\ z \in P_t(\epsilon_t) \text{ where } z_f := \mathbb{P}_t[u_f = 1],\ \forall f \in \mathcal{F}\right\}. \tag{EC40}$$

This model, which is inspired by a related model in (Agrawal et al. 2010), satisfies Assumption 2 as long as the confidence interval $P_t(\epsilon_t)$ is polyhedral; we formalize this next.

Proposition 9 (**Known Marginals**). *If* $\mathbb{P}_t = \mathcal{P}_t^{\text{corr}}$ *defined in* (EC40) *and* $P_t(\epsilon_t)$ *is a polytope,* $P_t(\epsilon_t) = \{q \in \mathbb{R}^{\mathcal{F}} : Aq \leq b, q \geq 0\}$ *for* $A \in \mathbb{R}^{\mathcal{F}\times K}$ *and* $b \in \mathbb{R}^K$*, the optimal profit* $\hat{\pi}_t(r)$ *in* (1) *can be obtained by solving the following convex optimization problem in variables* $y \in \mathbb{R}^K$*:*

$$\hat{\pi}_t(r) = \inf_{\theta, y\geq 0}\ \theta + y^\top b \tag{EC41a}$$

$$\text{s.t. } (p\cdot q - r - A^\top y)^\top d_t - c_t(d_t) \leq \theta, \quad \forall d_t \in \mathcal{I}_t \tag{EC41b}$$

Moreover, constraint (EC41b) *can be separated through oracle* (5)*.*

Proof. Consider the function $g(u) := \max_{d\in\mathcal{I}_t, d\leq u}\left((p\cdot q - r)^\top d - c_t(d)\right)$ defined for $u \in \{0,1\}^{\mathcal{F}}$ and let $p \in \mathbb{R}^{\{0,1\}^{\mathcal{F}}}$ denote a distribution from $\Delta_{\mathcal{F}}$. Intermediary t's profit from deviating can be written as:

$$\hat{\pi}_t(r) \ = \ \max_{p}\Big\{ \sum_{u\in\{0,1\}^{\mathcal{F}}} g(u)\, p_u \ :\ \sum_{u\in\{0,1\}^{\mathcal{F}}} p_u = 1,\ A\mu(p) \leq b,\ p \geq 0\Big\}, \tag{EC42}$$

where $\mu(p) \in \mathbb{R}^{\mathcal{F}}$ denotes the marginals of p, or equivalently, the mean of a random vector $\tilde{u} \in \{0,1\}^{\mathcal{F}}$ taking value u with probability p_u:

$$[\mu(p)]_f = \sum_{u\in\{0,1\}^{\mathcal{F}}} p_u \mathbf{1}(u_f = 1) \quad \Leftrightarrow \quad \mu(p) = \sum_{u\in\{0,1\}^{\mathcal{F}}} u\, p_u.$$

Consider $z \in \mathbb{R}$ as a new decision variable. We show that (EC42) has the same optimal value as the following program:

$$\max_{p,z}\Big\{ \sum_{u\in\{0,1\}^{\mathcal{F}}} g(u)\, p_u \ :\ \sum_{u\in\{0,1\}^{\mathcal{F}}} p_u = 1,\ Az \leq b,\ \mu(p) \leq z,\ p \geq 0\Big\}. \tag{EC43}$$

To prove this, notice first that (EC43) is a relaxation of (EC42) (if p is optimal for (EC42) and if $z = \mu(p)$, then (p, z) is feasible for (EC43) and yields the same objective). Now, consider (p, z) optimal for (EC43); we will construct p' which will yield a larger value for (EC42) than (p, z) does for (EC43). Let $\tilde{u}$ be the random vector taking value $u \in \{0,1\}^{\mathcal{F}}$ with probability p_u, and define a new random vector $\tilde{u}'$ as $\tilde{u}'_f = \tilde{u}_f + \tilde{e}_f \cdot \mathbf{1}(\tilde{u}_f = 0)$, where $\tilde{e}_f$ is a Bernoulli variable

taking value 1 with probability $(z_f - \mu(p)_f)/(1-\mu(p)_f)$ and value 0 with the complementary probability. Clearly, u' takes values in $\{0,1\}^{\mathcal{F}}$. Let p' be the distribution of $\tilde{u}'$ and let us calculate its marginals:

$$\begin{aligned}\mathbb{P}[\tilde{u}'_f = 1] &= \mathbb{P}[\tilde{u}_f = 1] + \mathbb{P}[\tilde{u}_f = 0] \cdot (z_f - \mu(p)_f)/(1-\mu(p)_f) \\ &= \mu(p)_f + (1-\mu(p)_f) \cdot (z_f - \mu(p)_f)/(1-\mu(p)_f) = z_f.\end{aligned}$$

This implies that $A\mu(p') \le b$ and therefore p' is feasible in (EC42). Moreover, $\mathbb{E}_{\tilde{u}'\sim p'}[g(\tilde{u}')] \ge \mathbb{E}_{\tilde{u}\sim p}[g(\tilde{u})]$ because $g(u)$ is increasing in u and $\tilde{u}' \ge \tilde{u}$ almost surely, giving the desired result.

This implies that $\hat{\pi}_t(r)$ is given by the optimal value of (EC43). Because (EC43) has a bounded and non-empty feasible set, its optimal value equals that of its dual, which is given by:

$$\min_{\theta\in\mathbb{R}, y\in\mathbb{R}^K_+, w\in\mathbb{R}^{\mathcal{F}}_+} (\theta + y^\top b) \quad \text{s.t.} \quad g(u) - \theta - \sum_{f\in\mathcal{F}} w_f u_f \le 0, \forall u \in \{0,1\}^{\mathcal{F}}, \quad w - A^T y \le 0. \tag{EC44a}$$

By changing the $u \in \{0,1\}^{\mathcal{F}}$ quantifiers for $d \in \mathcal{I}_t$, we obtained the desired result. □

References

Agrawal S, Ding Y, Saberi A, Ye Y (2010) Correlation robust stochastic optimization. *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, 1087–1096 (SIAM).

Archetti C, Speranza MG (2008) The split delivery vehicle routing problem: A survey. *The vehicle routing problem: Latest advances and new challenges*, 103–122 (Springer).

Bayraksan G, Love DK (2015) Data-driven stochastic programming using phi-divergences. *The Oper. Res. revolution*, 1–19 (INFORMS).

Ben-Tal A, Den Hertog D, De Waegenaere A, Melenberg B, Rennen G (2013) Robust solutions of optimization problems affected by uncertain probabilities. *Man. Sci.* 59(2):341–357.

Gounaris CE, Wiesemann W, Floudas CA (2013) The robust capacitated vehicle routing problem under demand uncertainty. *Oper. Res.* 61(3):677–693.

Kou L, Markowsky G, Berman L (1981) A fast algorithm for steiner trees. *Acta informatica* 15:141–145.

Schrijver A, et al. (2003) *Combinatorial optimization: polyhedra and efficiency*, volume 24 (Springer).

Stienen V, den Hertog D, Wagenaar J, de Zegher JF (2024) Enhancing digital road networks for better transportation in developing countries. *Transportation Research Interdisciplinary Perspectives* 27:101217.

Train KE (2009) *Discrete choice methods with simulation* (Cambridge university press).

truckmagz (2020) Pengaruh Rasio Konsumsi BBM Kendaraan terhadap Tarif Angkutan .

US Energy Information Administration (2022) Monthly Energy Review .

Yngwie Yudhistira W (2019) Pengukuran dinamis emisi karbondioksida (CO2) terhadap truk logistik untuk mendukung rantai pasok hijau .