

Many-to-many Matching with Cluster-Approximated Spatial Constraints*

Furkan Doğan[†]

Ömer Faruk Şahin[‡]

Özgür Yılmaz[§]

August 2026

Abstract

We study many-to-many matching with exact agent demands, exact slot quotas, and spatial composition constraints represented by a partition of agents into proximity clusters. At most one agent from each cluster may be assigned to a given slot. Although this local constraint is matroidal and induces a substitutable fixed-priority choice rule, exact demands and quotas couple completion globally across slots. Consequently, Deferred Acceptance and myopic Serial Dictatorship may terminate without a complete allocation even when one exists. In the quota-two benchmark, Deferred Acceptance has a uniform complete-feasibility guarantee if and only if there are exactly two clusters or the spatial constraint is nonbinding because every cluster is a singleton. We characterize completable partial allocations through cluster-based Hall inequalities and implement the resulting completion test using maximum flow. Building on this characterization, we introduce a completion-preserving priority mechanism that, under heterogeneous demands and capacities, produces a feasible and Pareto-efficient allocation satisfying replica-level no justified-envy under the layered priority order. It also satisfies replica-level pairwise stability when every slot has quota two. Simulations calibrated with administrative data on pharmacy duty schedules in Türkiye show that the mechanism clears the market in every reported scenario and improves model-implied rank outcomes under the generated preference profiles. These conclusions are robust to alternative cluster constructions, geographic specifications, and preference calibrations.

Keywords. Market Design, Scheduling, Spatial Constraints, Serial Dictatorship

JEL Classification Numbers: C78, D47, C61, D63

*Özgür Yılmaz acknowledges the research support from TÜBİTAK through Grant 223K459 and Furkan Doğan acknowledges the support from TÜBİTAK within the scope of the 2211-A National Doctoral Scholarship Program.

[†]Koç University, Department of Economics, İstanbul, 3445 Türkiye. Necmettin Erbakan University, Department of Economics, Konya, 42090 Türkiye. *E-mail address:* fdogan19@ku.edu.tr.

[‡]Stanford University, Department of Economics, Stanford, CA, 94305. *E-mail address:* osahin25@stanford.edu.

[§]Corresponding author. Koç University, Department of Economics, İstanbul, 34450 Türkiye. *E-mail address:* ozyilmaz@ku.edu.tr. Personal webpage: mysite.ku.edu.tr/ozyilmaz

1 Introduction

We study a shift-scheduling problem with fixed geographic service sites in which each shift must select a sufficiently dispersed set of locations to ensure equitable and convenient access. We focus on a tractable class of such problems in which spatial distribution requirements are approximated by proximity clusters. Agents within the same cluster are treated as too close to be assigned jointly, whereas agents from different clusters are treated as mutually compatible. We formalize this requirement as a zone-coverage constraint that restricts which agents may be assigned jointly. Combined with hard quotas on both sides of the market—each agent must work an exact number of shifts and each shift has fixed capacity—these spatial constraints severely limit the set of feasible allocations. In this environment, feasibility is not a peripheral technicality but the core design challenge: standard Deferred Acceptance (DA) and myopic Serial Dictatorship (SD) mechanisms can break down, leading to deadlock even when feasible completions exist, and conventional notions of efficiency, fairness, and stability become mutually incompatible with the geographic structure. Our central objective is therefore to characterize the feasible set in a way that enables implementable mechanisms and a coherent reformulation of the main axioms.

A scheduling problem under cluster-approximated spatial distribution constraints. Our analysis is motivated by a broad class of scheduling problems (see Section 5), most notably the assignment of pharmacy duty shifts in Türkiye, which serves as a running example throughout the paper. Duty shifts are determined annually at the district level, with heterogeneous capacities across days and holidays. Each pharmacy is entitled to a fixed number of shifts per year, and shift-level revenue coefficients provide an observable proxy for the relative attractiveness of duty shifts. To balance cumulative revenues, the regulating authority tracks realized earnings over multiple years and implicitly lowers the priority of pharmacies with higher past revenues. Spatial distribution is enforced through fixed zone-covering pairs—an institutional simplification of a more general separation constraint: pharmacies within a designated pair are typically assigned jointly to the same shifts throughout the year.

While institutionally functional, this centralized system exhibits three structural weaknesses. First,

it disregards pharmacies’ heterogeneous preferences over shifts, effectively assuming that revenue fully captures welfare. Second, by fixing zone-coverage to permanent pairs, it imposes rigid spatial constraints that block otherwise feasible reallocations. Third, assignments rely on an *ad hoc* revenue-equalizing heuristic lacking formal justification or transparency. These features collectively limit welfare and obscure the normative basis of allocation decisions. Consistent with the view that market design should remedy institutional flaws “[. . .] while adhering as closely as possible to current policy goals and the existing setup” (Sönmez, 2025), our objective is to provide a mechanism that preserves hard quotas and spatial requirements while restoring formal guarantees of feasibility, fairness, and efficiency.

Model overview and structural insight. We formulate a two-sided, many-to-many matching model with hard quotas and cluster-approximated spatial compatibility constraints. Agents (pharmacies) have preferences over slots (duty shifts), and all slots use a common priority ordering over agents. Each agent must be assigned exactly a fixed number of slots, and each slot has a fixed capacity.¹ Spatial constraints restrict which subsets of agents may be jointly assigned to a slot.

We model spatial requirements as a binary zone-coverage relation over agents. Under the maintained negative-transitivity restriction, this relation induces a partition of agents into proximity clusters (Lemma 1). Agents within the same cluster are too geographically close to be assigned jointly, whereas agents from distinct clusters form admissible pairs. This partition structure is the key to understanding feasibility. When geographic constraints involve overlapping neighborhoods or slot-specific admissibility, the partition representation no longer applies, and the exact characterization results of this paper need not hold in the same form.

The induced geometry yields an exact characterization of DA’s uniform feasibility guarantee. In the single-unit-demand environment with homogeneous slot quota two, DA produces a feasible allocation for every common priority ordering and preference profile if and only if either the zone-coverage relation induces exactly two clusters or every cluster is a singleton (Theorem 1). The latter is a non-binding case: when all clusters are singletons, every pair of distinct agents is compatible and the spatial

¹In our application, each pharmacy is guaranteed a fixed number of duty shifts within a calendar year; accordingly, institutional practice treats these requirements as hard quotas.

constraint imposes no effective restriction.

Conditional on a non-vacuous zone-coverage constraint, the two-cluster case is therefore the exact boundary for DA’s uniform feasibility guarantee. With two clusters and quota two, every full feasible slot contains one agent from each cluster, so completing a slot preserves the cluster balance required by the residual problem. When the constraint is non-vacuous and there are three or more clusters, a locally admissible slot may be filled using only a subset of clusters. The resulting exclusion may leave residual cluster demands and residual slot capacities that cannot be jointly completed. Thus, the relevant distinction is between local admissibility and global completion: the former need not preserve the latter.

Mechanism design under global feasibility. To address this structural failure, we introduce Feasible Serial Dictatorship (*FSD*) for the single-unit-demand case, a mechanism that assigns agents sequentially while enforcing global completion at every step. Whenever a feasible allocation exists, *FSD* produces one consistent with the priority order (Theorem 3). By embedding a forward-looking feasibility test into each assignment decision, the mechanism eliminates structural dead-ends. In this sense, *FSD* is a completion-preserving version of serial dictatorship: it accepts a tentative assignment only when the resulting partial allocation remains extendable to a feasible full allocation.

We extend the analysis to multi-unit demand environments. The resulting Generalized Feasible Serial Dictatorship (*GFSD*) ensures feasibility through a generalized completion condition characterized using tools from network flow theory, specifically the Ford–Fulkerson Theorem and the Integrality Theorem (Schrijver, 2003). This characterization (Theorem 4) translates the geometric constraint into a global flow condition that is both necessary and sufficient for implementability. *GFSD* uses this condition dynamically: each tentative assignment is accepted only if the residual problem still admits a feasible completion.

Normative properties. The allocations produced by *FSD* and *GFSD* satisfy framework-appropriate versions of no justified envy and Pareto efficiency under spatial constraints (Theorems 3 and 5). Pairwise stability is established in the quota-two environment. That is, for *GFSD*, this result allows heterogeneous demands but for each $s \in \mathcal{S}$, requires $q_s = 2$. For *GFSD*, no justified-envy protects

individual replicas under the layered priority order $\succ^\Gamma$, not original agents’ complete bundles under $\succ$; Appendix D provides a counterexample to the latter property. We also clarify the mechanisms’ incentive properties: *FSD* is strategy-proof in the single-unit environment, whereas *GFSD* is generally manipulable under multi-unit demand. Where necessary for the normative analysis, we extend preferences over slots to bundles using a parsimonious ordinal lexicographic rule that requires neither cardinal utilities nor additional assumptions about complementarities across shifts.

Accordingly, the empirical analysis is interpreted as a counterfactual comparison under preferences generated by the maintained Data-Generating Process (DGP), rather than as an equilibrium prediction under strategic reporting. These results show that strong normative properties can be restored while retaining the maintained demand, capacity, and cluster-composition constraints; they also identify the incentive limits of sequential completion-preserving allocation in the multi-unit domain.

Summary of contributions. This paper makes three main contributions. First, we introduce a matching model with hard quotas and spatial incompatibility constraints and provide a structural characterization of feasibility: under the maintained negative-transitivity restriction, the zone-coverage relation induces proximity clusters that determine exactly when allocations can be completed. Second, we identify a sharp threshold for *DA*—it has a uniform feasibility guarantee if and only if either there are exactly two clusters or the spatial constraint is nonbinding because every cluster is a singleton—and show that with three or more clusters classical local acceptance logic can generate inherent dead-ends. This isolates a precise boundary at which standard local acceptance logic no longer preserves feasible completion. We further show in the appendix that this threshold is the quota-two instance of a broader capacity–cluster balance. Third, we characterize the conditions under which partial allocations admit global completion and design a mechanism that ensures global completability at every step. We show that this mechanism restores feasibility while preserving the normative properties described above. Together, these results provide a tractable framework for mechanism design in environments where geographic structure fundamentally constrains implementability.

Empirical validation. We conclude by simulating the pharmacy scheduling environment using administrative data from Türkiye. Spatial constraints are represented by partitions recovered from histor-

ical co-duty assignments and geographic compactness. Preferences are modeled using revenue-based attractiveness with noise, and priorities reflect historical revenue equalization. We compare the current heuristic mechanism with algorithmic benchmarks. Across the reconstructed environments, the locally partner-restricted heuristic and myopic allocation procedure may fail to clear the market or produce poor model-implied rank outcomes. *GFSD* clears the market in every reported scenario and improves model-implied rank outcomes under the simulated preference profiles. The qualitative feasibility ranking is stable across the threshold, geographic, and preference specifications considered.

Relation to the literature. We contribute to market design by studying priority-based allocation under rich feasibility constraints with a spatial structure. Our work is related to the canonical literature on matching with contracts, many-to-many stability, and constrained allocation. Matching with contracts provides a general framework in which substitutability and related conditions support stable outcomes and incentive properties ([Hatfield and Milgrom, 2005](#)). Many-to-many matching raises additional stability issues, and substitutability-type restrictions play a central role in ensuring the existence and structure of stable outcomes ([Echenique and Oviedo, 2006](#)). A significant strand of the literature examines whether classical rules—Deferred Acceptance (DA) and Serial Dictatorship (SD)—remain well-behaved under distributional or capacity constraints. The dominant paradigm identifies substitutability-type conditions ([Hatfield and Kojima, 2010](#))—often formalized through hierarchical (laminar) caps ([Kamada and Kojima, 2018](#)), matroidal structures ([Kojima, Tamura, and Yokoo, 2018](#)), or polyhedral quotas ([Kojima, Sun, and Yu, 2020](#))—under which decentralized acceptance procedures retain desirable stability, incentive, or feasibility properties.

The appropriate comparison with this literature requires distinguishing local feasibility from global completion. The connection to matroidal and laminar structures is exact at the local level. At an individual slot, the proximity clusters are pairwise-disjoint subsets of the common ground set of agents. The restriction that at most one agent may be selected from each cluster defines a partition matroid, truncated by the slot’s overall capacity. Equivalently, the locally admissible agent sets form a laminar matroid ([Barman and Biswas, 2020](#); [Fife and Oxley, 2017](#)). The associated fixed-priority greedy choice is therefore substitutable ([Yokoi, 2019](#)); in the quota-two benchmark, Proposition 1 establishes

this property directly. Our environment therefore does not depart from the matching literature through a failure of local matroidality or local substitutability.

The additional issue is complete feasibility under exact requirements on both sides. Local substitutability disciplines acceptance decisions at an individual slot, but it does not imply that independently admissible slot holdings can be extended to an allocation satisfying every agent’s exact demand and every slot’s exact quota. Assigning an agent to a slot uses not only ordinary slot capacity but also the cluster-specific opportunity at that slot: once a cluster is represented, no additional agent from that cluster can be assigned there. Consequently, a collection of locally admissible slot holdings may leave residual cluster demands and residual capacities that cannot be jointly completed. The relevant object is therefore not merely a collection of independent local choices, but the set of partial allocations that admit a feasible completion. This is why *DA* may terminate without a complete feasible allocation even though every slot uses a substitutable choice function and a complete feasible allocation exists.

This global completion problem nevertheless has a precise matroidal structure. For each partial allocation, each slot induces a residual laminar matroid describing which remaining units of demand can still be assigned there. A completion exists exactly when all remaining demand units can be distributed across these slot-specific matroids; formally, this is a matroid-union problem. Their union is itself a matroid, but it need not be laminar and changes with the partial allocation because each tentative assignment alters both residual capacity and cluster availability. Edmonds’s matroid-union theorem provides a rank characterization of completion ([Edmonds, 1968](#)). [Appendix C](#) shows that these rank inequalities reduce exactly to the cluster-based Hall condition in [Theorem 4](#), thereby providing a matroidal interpretation of the maximum-flow characterization underlying *GFSD*.

Our results are therefore complementary to the substitutability and matroidal-constraint approaches. The novelty lies not in introducing a non-substitutable local choice environment, but in characterizing and dynamically preserving exact global completion. [Theorem 1](#) identifies the capacity–cluster boundary at which *DA* preserves completion, while [Theorem 4](#) characterizes the partial allocations that remain completable. *FSD* and *GFSD* differ from myopic local procedures by enforcing the corresponding completion conditions dynamically, using a polynomial-time completion oracle based on

maximum flow in the general environment.

Our paper is also related to the literature on distributional constraints, controlled choice, and hard quotas. In regional matching and controlled school choice, constraints typically regulate the number or type distribution of agents assigned across regions, schools, or institutions ([Kamada and Kojima, 2015](#); [Echenique and Yenmez, 2015](#)). Hard distributional constraints and dynamic quota mechanisms have been studied as a way to improve upon artificial-cap approaches while preserving desirable fairness and incentive properties ([Fragiadakis and Troyan, 2017](#)). The constraint studied here has a different operational structure: it is per-slot and compositional, so that no slot may contain two agents from the same cluster. Once this per-slot composition rule is combined with hard agent demands and fixed slot capacities, feasibility becomes globally coupled across the entire schedule. Relatedly, daycare allocation models impose ratio constraints through global caps ([Kamada and Kojima, 2024](#)). Unlike adjustable cutoffs in those settings, our framework maintains fixed priorities and preserves feasible completion through a completion characterization.

Closely related work studies serial allocation under general feasibility constraints. [Imamura and Kawase \(2024\)](#), for example, formulate constrained serial dictatorship on a fixed family of feasible matchings, while [Imamura and Kawase \(2025\)](#) study efficiency and strategy-proofness under general constraints and relate their results to constrained serial dictatorship. *GFSD* uses the same broad constrained serial allocation logic but supplies a domain-specific completion oracle. In such settings, myopic greedy assignments can generate structural dead-ends, requiring feasibility to be enforced dynamically through a global completion condition.²

Multi-unit demand introduces further complications, as sequential priority assignment may conflict with bundled preferences ([Manea, 2007](#)). We address this by reducing multi-unit requests to single-unit choices through a replica-market construction while maintaining priority consistency, allowing us to focus on global feasibility under spatial constraints.

²Similar breakdowns of myopic allocation logic arise in seat allocation with gender-based restrictions ([Afacan, Dur, and Dur, 2025](#)), where adaptive procedures are introduced to preserve feasibility under additional constraints.

2 Model

We study a scheduling problem in which a finite set of slots $\mathcal{S}$ are allocated to a finite set of agents $\mathcal{A}$. Each agent $i \in \mathcal{A}$ has demand $d_i \in \mathbb{Z}_+$, and each slot $s \in \mathcal{S}$ has capacity $q_s \in \mathbb{Z}_+$, with $\sum_{i \in \mathcal{A}} d_i = \sum_{s \in \mathcal{S}} q_s$. Let $d = (d_i)_{i \in \mathcal{A}}$ and $q = (q_s)_{s \in \mathcal{S}}$ denote the demand and capacity vectors.

Each $i \in \mathcal{A}$ has a strict preference relation π_i over $\mathcal{S}$, with weak and strict parts π_i^R and π_i^P . The preference profile is $\pi = (\pi_i)_{i \in \mathcal{A}}$. A *priority ordering*, applied at every slot, is a linear order $\succ$ on $\mathcal{A}$, with weak part $\succeq$, where $i \succeq j$ if and only if $i \succ j$ or $i = j$.

Two agents form a *zone-covering pair* if they are sufficiently distant to ensure adequate service-area coverage; such agents may therefore be assigned to the same slot. A slot must be covered by a set of agents in which every pair is zone-covering. We assume that zone-coverage is slot-independent: for any $s, s' \in \mathcal{S}$, agents i, j are zone-covering for s if and only if they are zone-covering for s' . This structure is represented by a symmetric and irreflexive binary relation $\mathcal{F} \subseteq \{(i, j) \in \mathcal{A} \times \mathcal{A} : i \neq j\}$, where $(i, j) \in \mathcal{F}$ if and only if i and j form a zone-covering pair. We refer to $\mathcal{F}$ as the *zone-coverage relation*, and call j a *zone-covering partner* of i whenever $(i, j) \in \mathcal{F}$. For any $A \subseteq \mathcal{A}$, let $\mathcal{F}_A = \{(i, j) \in A \times A : (i, j) \in \mathcal{F}\}$ denote the set of zone-covering pairs in A .

A *scheduling problem under spatial distribution constraints* is specified by $(\mathcal{A}, \mathcal{S}, \mathcal{F}, d, q, \pi, \succ)$. An *allocation* is a function $\mu : \mathcal{A} \rightarrow \mathcal{P}(\mathcal{S})$ assigning each agent a set of slots, with $\mu(i)$ denoting agent i 's assignment. For each slot $s \in \mathcal{S}$, let $\mu_s = \{i \in \mathcal{A} : s \in \mu(i)\}$ denote the agents assigned to s . Fixing $(\mathcal{A}, \mathcal{S}, \mathcal{F}, d, q, \succ)$, a *mechanism* φ maps each reported preference profile π to an allocation $\mu^\varphi(\pi)$, with $\mu_i^\varphi(\pi)$ denoting agent i 's assignment.

Definition 1 An allocation μ is *feasible* if,

1. For each $i \in \mathcal{A}$, $|\mu(i)| = d_i$ (demand condition)
2. for each $s \in \mathcal{S}$, $|\mu_s| = q_s$ (quota condition)
3. for each $s \in \mathcal{S}$ and all distinct $i, j \in \mu_s$, $(i, j) \in \mathcal{F}$ (zone-coverage condition).

2.1 Structure of the spatial distribution constraints

We approximate $\mathcal{F}$ by imposing the substantive restriction that whenever i and j are too close to form a zone-covering pair and j and k are too close as well, then i and k are also deemed too close. Formally, $(i, j) \notin \mathcal{F}$ and $(j, k) \notin \mathcal{F}$ imply $(i, k) \notin \mathcal{F}$, hence $\mathcal{F}$ is *negatively transitive*. Under the natural symmetry of “too close,” this is equivalent to a partition of agents into proximity clusters: agents within the same cluster are geographically too close to share a slot (all within-cluster pairs lie in the complement relation of $\mathcal{F}$), whereas agents from different clusters form zone-covering pairs (all cross-cluster pairs lie in $\mathcal{F}$).

Lemma 1 *Let $\mathcal{F}$ be an irreflexive binary relation on $\mathcal{A}$. $\mathcal{F}$ is symmetric and negatively transitive if and only if there exists a partition $\{A_1, \dots, A_m\}$ of $\mathcal{A}$ such that $(i, j) \in \mathcal{F}$ if and only if i and j belong to different sets in this partition.*

To see the induced structure, define $i \sim j$ if and only if $i = j$ or $(i, j) \notin \mathcal{F}$. Under symmetry and negative transitivity, $\sim$ is an equivalence relation, and its equivalence classes are the induced proximity clusters. Hence, the zone-coverage graph is complete multipartite: within-cluster pairs are inadmissible and all cross-cluster pairs are admissible.

We note that the partition structure characterized in Lemma 1 need not represent all zone-coverage constraints implied by a real geographical topology, which may induce intersecting clusters or vary across slots. In such environments, the partition representation and the resulting exact characterizations need not extend. Nevertheless, modeling zone-coverage as a negatively transitive binary relation $\mathcal{F}$ provides a tractable partition approximation.

We refer to the resulting partition $\{A_1, \dots, A_m\}$ as the *(induced) clusters*. Whenever convenient, we refer to a generic cluster A_ℓ simply by its index ℓ . The cluster index of agent i is denoted by $[i]$; that is, $i \in A_{[i]}$. Throughout, $\mathcal{F}$ and its induced clusters are fixed.

Assumption 1 *The zone-coverage relation $\mathcal{F}$ is symmetric and negatively transitive, and $\{A_1, \dots, A_m\}$ are the induced clusters.*

Whether the induced cluster structure admits a feasible allocation depends, in particular, on the capacity vector $q = (q_s)_{s \in \mathcal{S}}$. Feasibility immediately implies $q_s \leq m$ for each slot s . We refer to the subclass induced by Assumption 1 as *scheduling problems under cluster-approximated spatial distribution constraints*, or simply as *problems* hereafter.

3 Many-to-many Matching under Hard Quotas and Spatial Distribution Constraints

We begin with the special case of single-unit demands (Section 3.1), which highlights the main challenge and develops the core intuition for our analysis. We then extend the analysis to multi-unit demands and heterogeneous capacities (Section 3.2). While this general model builds on the same intuition, it introduces additional technical challenges, which we address in the Appendix.

3.1 Single-unit-demand Problem

We consider the simplest version of the problem. For spatial distribution constraints to bind, slots must have capacity at least two. Accordingly, we begin with the single-unit-demand case, in which each agent $i \in \mathcal{A}$ has demand $d_i = 1$ and each slot $s \in \mathcal{S}$ has capacity $q_s = 2$. The quota condition then implies $|\mathcal{A}| = 2|\mathcal{S}|$. We identify each singleton assignment with its unique slot.

3.1.1 Axioms. We first introduce the standard notion of efficiency.

Definition 2 A feasible allocation μ is **Pareto efficient** if there exists no feasible allocation μ' such that for each $i \in \mathcal{A}$, $\mu'(i) \pi_i^R \mu(i)$ and for at least one $j \in \mathcal{A}$, $\mu'(j) \pi_j^P \mu(j)$.

We consider a fairness requirement that rules out justified envy. An allocation satisfies *no justified-envy* with respect to $\succ$ if a higher-priority agent who prefers a lower-priority agent's assignment could obtain that bundle only by harming an agent ranked above them. This property provides strict top-down protection of priority claims.

Definition 3 An allocation μ satisfies **no justified-envy** under $\succ$ if for any $i \succ j$ with $\mu(j) \pi_i^P \mu(i)$, and any feasible μ' with $\mu'(i) = \mu(j)$, there exists $k \succ i$ such that $\mu(k) \pi_k^P \mu'(k)$.

We next introduce a stability notion based on pairwise deviations. To this end, we extend the agents' priority ordering to pairs. Specifically, we represent any two-agent set $\{i, j\}$ by the ordered tuple $(i, j)_{\succ}$, where $i \succ j$. We then define an ordering $\succeq_c$ over agent pairs as

$$(i, j)_{\succ} \succeq_c (i', j')_{\succ} \text{ if and only if } i \succeq i' \text{ and } j \succeq j',$$

with strict part $(i, j)_{\succ} \succ_c (i', j')_{\succ}$ whenever at least one inequality is strict and $(i, j) \neq (i', j')$. We refer to this extension as *pair-priority ordering*.

Given a feasible allocation μ , a zone-covering pair (i, j) *blocks* μ if there exists $s \in \mathcal{S}$ such that $s \pi_i^P \mu(i)$ and $s \pi_j^P \mu(j)$, $(i, j) \in \mathcal{F}$, and $(i, j) \succ_c \mu_s$. A zone-covering pair (i, j) *weakly blocks* μ if there exists $s \in \mathcal{S}$ such that $s \pi_i^R \mu(i)$ and $s \pi_j^R \mu(j)$, at least one of these preference comparisons is strict, $(i, j) \in \mathcal{F}$, and $(i, j) \succeq_c \mu_s$. The notion of stability follows from this definition of a two-agent deviation at a single slot.

Definition 4 A feasible allocation μ is *pairwise* (resp. *strongly*) *stable* if no zone-covering pair (i, j) (resp. *weakly*) blocks μ .

The definition is pair-based since the minimal feasible assignment to a slot is not a single agent but a zone-covering pair. A slot cannot evaluate an agent independently of her potential partner: even a high-priority agent can be assigned to a slot only together with another agent who satisfies the zone-coverage relation. Thus, the relevant local deviation is not that an individual agent joins a slot, but that an admissible pair of agents jointly replaces the current pair at that slot. Pairwise stability should therefore be read as the absence of locally admissible zone-covering-pair deviations. The notion is local and spatially constrained: it checks whether a single slot can be profitably recomposed by a zone-covering pair. A blocking deviation is not required to satisfy a separate global completion test. Adding such a requirement would restrict the set of blocking deviations and therefore yield a weaker stability notion; global completion is instead enforced by the mechanisms introduced below.

We finally define the incentive property used in the mechanism analysis.

Definition 5 A mechanism φ is **strategy-proof** if there exist no agent $i \in \mathcal{A}$, preference profile $\pi = (\pi_i, \pi_{-i})$, and alternative report $\widehat{\pi}_i$ such that

$$\mu_i^\varphi(\widehat{\pi}_i, \pi_{-i}) \pi_i^P \mu_i^\varphi(\pi_i, \pi_{-i}).$$

3.1.2 Deferred Acceptance. We next analyze the well-known *deferred acceptance (DA)*. For each $A \subseteq \mathcal{A}$, define $F_A = \{(i, j)_\succ \in A \times A : (i, j) \in \mathcal{F}\}$.

Lemma 2 For each $A \subseteq \mathcal{A}$, either $F_A = \emptyset$, or there exists $(i_A^*, j_A^*)_\succ \in F_A$ such that for each $(i, j)_\succ \in F_A$, $(i_A^*, j_A^*)_\succ \succeq_c (i, j)_\succ$. Moreover, in the latter, i_A^* is the $\succ$ -maximal agent in A .

The pair-priority ordering induces a canonical choice from any agent set in $\mathcal{A}$. For each $A \subseteq \mathcal{A}$, if $F_A \neq \emptyset$, then by Lemma 2, (i_A^*, j_A^*) exists and i_A^* is the $\succ$ -maximal agent in A . For nonempty $A \subseteq \mathcal{A}$ with $F_A = \emptyset$, let i_A^* denote the $\succ$ -maximal agent in A . We define the (*priority-respecting*) *choice function* $C_s: 2^{\mathcal{A}} \rightarrow 2^{\mathcal{A}}$ as follows:

$$C_s(A) = \begin{cases} \emptyset, & \text{if } A = \emptyset, \\ \{i_A^*, j_A^*\}, & \text{if } F_A \neq \emptyset, \\ \{i_A^*\}, & \text{if } A \neq \emptyset \text{ and } F_A = \emptyset. \end{cases}$$

The transformation from an incomplete pair order to a choice function allows us to analyze how properties of $\mathcal{F}$ develop into desirable properties of C_s . We say, a choice function $C: 2^{\mathcal{A}} \rightarrow 2^{\mathcal{A}}$ is *substitutable* if for each $A \subseteq B \subseteq \mathcal{A}$ and $i \in A$, whenever $i \in C(B)$, then $i \in C(A)$.

Proposition 1 For every symmetric and negatively transitive zone-coverage relation $\mathcal{F}$ and every priority ordering $\succ$, the priority-respecting choice function is single-valued and substitutable.

The result has a matroidal interpretation. More generally, for each slot s with capacity q_s , let

$$\mathcal{I}_s = \{I \subseteq \mathcal{A} : |I| \leq q_s \text{ and } |I \cap A_\ell| \leq 1 \text{ for every } \ell = 1, \dots, m\}$$

denote the family of locally admissible agent sets. Ignoring the overall capacity constraint, the restriction that at most one agent may be selected from each proximity cluster defines a partition matroid whose blocks are the clusters and whose block capacities are one. Imposing the additional bound $|I| \leq q_s$ gives its truncation at rank q_s . Equivalently, $(\mathcal{A}, \mathcal{I}_s)$ is the laminar matroid generated by the family $\{\mathcal{A}, A_1, \dots, A_m\}$, with capacity q_s on $\mathcal{A}$ and capacity one on each cluster A_ℓ (Fife and Oxley, 2017).

In the quota-two environment considered here, the priority-respecting choice function defined above coincides with the fixed-priority greedy choice induced by this matroid: agents are considered in priority order and retained whenever their cluster is not yet represented, until the slot reaches capacity or all applicants have been considered. Because the priority ordering is strict, this greedy choice is single-valued. Moreover, fixed-priority greedy choice under a matroid constraint is substitutable (Yokoi, 2019). We nevertheless provide a short, self-contained proof of Proposition 1 in the Appendix.

Single-valuedness makes each slot's tentative holding unique, while substitutability places the slot-level choices within the standard substitutable-choice domain used by DA .

Deferred Acceptance under priority ordering $\succ (DA)$.

Step 1 Each agent proposes to her most preferred slot. Each slot s considers the set $A_s \subseteq \mathcal{A}$ of proposers it has received and tentatively holds $C_s(A_s)$, rejecting all others.

Step k Each agent rejected at Step $k - 1$, who has an untried slot, proposes to her most preferred remaining slot. Each slot considers new proposals and stand-by agents from the previous steps, applies C_s to this combined set, and rejects others.

The process ends when no rejections occur. Each agent is assigned to the slot that holds it, and each slot holds exactly the set of agents chosen in the final step by C_s .

Substitutability of choice functions (Proposition 1) ensures that DA terminates in finitely many rounds. However, as the next example shows, the DA may not result in a feasible outcome even if there exists one.

Example 1 Let $\mathcal{A} = \{i_1, i_2, i_3, i_4\}$ and $\mathcal{S} = \{s_1, s_2\}$. The set of zone-covering pairs is induced by the partition $\{A_1, A_2, A_3\} = (\{i_1, i_2\}, \{i_3\}, \{i_4\})$ such that $\mathcal{F} = \{(i, j) \in \mathcal{A} \times \mathcal{A} : [i] \neq [j]\}$. The priority ordering is given as $i_3 \succ i_4 \succ i_1 \succ i_2$. Let agents' preferences be $i_1: s_2 \pi_{i_1}^P s_1$, $i_2: s_2 \pi_{i_2}^P s_1$, $i_3: s_1 \pi_{i_3}^P s_2$ and $i_4: s_1 \pi_{i_4}^P s_2$. Observe that the following two allocations are feasible: in the first, agents $\{i_1, i_3\}$ and $\{i_2, i_4\}$ are assigned to slots s_1 and s_2 , respectively; in the second, agents $\{i_1, i_4\}$ and $\{i_2, i_3\}$ are assigned to slots s_1 and s_2 , respectively. However, the outcome of the DA, namely the allocation μ such that $\mu_{s_1} = \{i_3, i_4\}$, and $\mu_{s_2} = \{i_1\}$, violates feasibility.

Notice that each slot's holding is independent in its local matroid: $\{i_3, i_4\} \in \mathcal{I}_{s_1}$ and $\{i_1\} \in \mathcal{I}_{s_2}$. The failure is therefore not local. Agent i_2 is the only unmatched agent and belongs to the same cluster as i_1 , which is already represented at the only slot with residual capacity. Thus, for the singleton cluster set $L = \{1\}$, residual demand is one while residual effective capacity is zero.

This example motivates the following characterization for the DA's outcome being feasible.

Theorem 1 Let $\mathcal{F}$ be symmetric and negatively transitive and let $\{A_1, \dots, A_m\}$ be the induced proximity clusters. Suppose a feasible allocation exists. Then, for each $\succ$ and $(\pi_i)_{i \in \mathcal{A}}$, the DA outcome is feasible if and only if one of the following is true:

1. $m = 2$,
2. for each $\ell = 1, \dots, m$, $|A_\ell| = 1$.

Theorem 1 identifies two sources of DA's uniform feasibility guarantee. In the balanced case, $m = q = 2$, every full feasible slot contains one agent from each cluster, so completing a slot preserves the cluster balance required for completion. In the all-singleton case, the spatial constraint is non-binding because every pair of distinct agents is compatible. Apart from this all-compatible case, two clusters are necessary and sufficient for the uniform guarantee. When the constraint is non-vacuous and $m > 2$, locally admissible acceptances may alter residual cluster demands so that no feasible completion remains; DA does not test for this possibility. Appendix B extends the characterization to homogeneous quota q .

This motivates a mechanism with per-step global checks: the next subsection introduces a rule that accepts a tentative assignment only if the resulting partial allocation still admits a feasible completion.

3.1.3 Feasible Serial Dictatorship. We propose a version of the serial dictatorship mechanism that incorporates feasibility constraints. To this end, we introduce the notion of a partial allocation. A partial allocation specifies the assignments made up to a given step of the mechanism, subject to the requirement that all quota and zone-coverage conditions are respected.

Definition 6 *An allocation μ is **partial** if*

1. *For each $i \in \mathcal{A}$, $|\mu(i)| \leq 1$,*
2. *for each $s \in \mathcal{S}$, $|\mu_s| \leq 2$,*
3. *for each $s \in \mathcal{S}$ and all distinct $i, j \in \mu_s$, $(i, j) \in \mathcal{F}$.*

A partial allocation μ *admits a completion* if there exists a feasible allocation μ' such that for each $i \in \mathcal{A}$, $\mu(i) \subseteq \mu'(i)$. Each such allocation μ' is called a *completion*.

Let μ be an allocation. If $|\mu_s| = 1$, then we denote the unique agent in μ_s as $j_\mu(s)$. We denote the set of *vacant slots* by $S_0^\mu = \{s \in \mathcal{S} : |\mu_s| = 0\}$; for each $i \in \mathcal{A}$, the set of *partially filled slots* available for i by $S_1^\mu(i) = \{s \in \mathcal{S} : |\mu_s| = 1 \text{ and } (i, j_\mu(s)) \in \mathcal{F}\}$; the set of available slots for i by $S^\mu(i) = S_0^\mu \cup S_1^\mu(i)$; the set of available slots for agents in A by $S_A^\mu = \bigcup_{i \in A} S^\mu(i)$; and the set of agents in A_ℓ without an assigned slot under μ by $A_\ell(\mu) = \{i \in A_\ell : \mu(i) = \emptyset\}$.

A necessary condition for μ to admit a completion is $|A_\ell(\mu)| \leq |S_{A_\ell}^\mu|$ for every cluster ℓ (pigeon-hole principle): otherwise, some two agents in A_ℓ must share a slot in any completion, violating the zone-coverage condition in Definition 1. The converse also holds:

Theorem 2 *A partial allocation μ admits a completion if and only if for each $\ell = 1, \dots, m$, $|A_\ell(\mu)| \leq |S_{A_\ell}^\mu|$.*

We observe that under the empty allocation (no slots assigned), Theorem 2 reduces to $|A_\ell| \leq |\mathcal{S}|$ for each ℓ . Then, the quota condition ($|\mathcal{A}| = 2|\mathcal{S}|$) implies that each cluster contains at most half of all agents. We adopt this as a standing assumption for the remainder of the section.

For any partial allocation μ , we define $D_\ell(\mu) = |S_{A_\ell}^\mu| - |A_\ell(\mu)|$. By Theorem 2, there exists a feasible allocation if and only if for each ℓ , $D_\ell(\mu^0) \geq 0$. We define the set of *feasibly available* slots for i as follows:

$$S_{f-av}^\mu(i) = S_0^\mu \cup \left\{ s \in S_1^\mu(i) : \text{for each } \ell \notin \{[i], [j_\mu(s)]\}, D_\ell(\mu) > 0 \right\}.$$

We are now ready to introduce the serial dictatorship mechanism adjusted for feasibility, which refines the standard procedure to ensure that quota and zone-coverage constraints are respected.

Feasible Serial Dictatorship (FSD).

Step 0 Let μ^0 be the empty allocation.³

Step k Assign i_k to the π_{i_k} maximal element of $S_{f-av}^{\mu^{k-1}}(i_k)$. Denote the updated allocation by μ^k .

The mechanism *FSD* terminates at Step $k = |\mathcal{A}|$, and its outcome is the allocation $\mu^{|\mathcal{A}|}$. Our main result in this section shows that $\mu^{|\mathcal{A}|}$ is feasible, and satisfies no justified-envy, pairwise stability, and Pareto efficiency.

Theorem 3 For each $\succ$, *FSD* is strategy-proof. Moreover, for each $(\pi_i)_{i \in \mathcal{A}}$, the outcome of *FSD* satisfies feasibility, no justified-envy, pairwise stability and Pareto efficiency.

3.2 Multi-unit-demand Problem with Heterogeneous Capacities

We extend the single-unit-demand model of Section 3.1 to a setting with heterogeneous multi-unit demands and heterogeneous slot capacities. The notion of partial allocation carries over: allocations must respect slot capacities q_s while satisfying the zone-coverage condition.

Construction of replica agents. The construction has two steps: we first expand each agent into unit-demand replicas and then order those replicas consistently with the original priority order. For each agent i , we create $|d_i|$ replicas. This construction is essential: once agents have multi-unit demands, the extension of the priority ordering $\succ$ directly affects the interpretation of the axioms. For each $i \in \mathcal{A}$, we construct the set of replicas of i by $\Gamma(i) = \{i^1, i^2, \dots, i^{d_i}\}$, and for each $A \subseteq \mathcal{A}$, the set of

³Note that, as mentioned earlier, we impose that for each ℓ , $D_\ell(\mu^0) \geq 0$ such that the quota condition is satisfied.

all replicas of agents in A is denoted by $A^\Gamma = \bigcup_{i \in A} \Gamma(i)$. We endow $\mathcal{A}^\Gamma$ with the replica order $\succ^\Gamma$, defined by:

$$i^k \succ^\Gamma j^\ell \text{ if and only if } k < \ell \text{ or } (k = \ell \text{ and } i \succ j).$$

In the replica order $\succ^\Gamma$, (i) each agent's first replica precedes each agent's second replica, each second replica precedes each third replica, and so forth, and (ii) within each layer of replicas with the same index k , that is, the k th replicas of all agents, the ordering is identical to the priority ordering $\succ$. For example, consider three agents i, j, k , each demanding two units, with priority ordering $i \succ j \succ k$. Under our construction, the induced replica order $\succ^\Gamma$ is given by the following. $i^1 \succ^\Gamma j^1 \succ^\Gamma k^1 \succ^\Gamma i^2 \succ^\Gamma j^2 \succ^\Gamma k^2$.⁴

We choose the replica order $\succ^\Gamma$ to satisfy two minimal consistency requirements: (i) respect for $\succ$ —a lower-priority agent cannot overtake a higher-priority one merely because they demand fewer units; and (ii) symmetry across units—no agent's first unit is disadvantaged relative to another's later unit. Ordering replicas lexicographically by (replica index, then $\succ$) satisfies both: all agents compete on equal footing for their first unit before any second unit is considered, preventing high-demand agents from crowding out low-demand agents early in the allocation.

We extend the zone-coverage relation to replicas by defining $\mathcal{F}^\Gamma \subseteq \mathcal{A}^\Gamma \times \mathcal{A}^\Gamma$ such that:

$$(i^k, j^\ell) \in \mathcal{F}^\Gamma \text{ if and only if } i \neq j \text{ and } (i, j) \in \mathcal{F}.$$

Observe that if $\mathcal{F}$ is negatively transitive, so is $\mathcal{F}^\Gamma$; the induced replica clusters are $A_1^\Gamma, \dots, A_m^\Gamma$, where A_ℓ^Γ contains all replicas of agents in A_ℓ .

Residual demand. Let μ be a partial allocation. The set of replicas of agents in A_ℓ without an assigned slot under μ is denoted by $A_\ell^\Gamma(\mu) = \{i^k \in A_\ell^\Gamma : [i] = \ell \text{ and } \mu(i^k) = \emptyset\}$. Then, for each $L \subseteq \{1, \dots, m\}$, the total residual demand is $\sum_{\ell \in L} |A_\ell^\Gamma(\mu)|$.

Residual supply. The availability of a slot to an agent in cluster ℓ is constrained by both its remaining

⁴Note that an alternative replica ordering, say $\succ^\Delta$, could be that $i^1 \succ^\Delta i^2 \succ^\Delta j^1 \succ^\Delta j^2 \succ^\Delta k^1 \succ^\Delta k^2$. Under such an order, an agent with larger demand may exhaust available capacity before lower-priority agents receive any assignment, even when a feasible allocation exists that assigns at least one unit to each agent.

capacity and the clusters of the agents assigned to the slot. Let $q_s^\mu = q_s - |\mu_s|$ be the residual capacity of s at μ . For each s , let $L_s^\mu \subseteq \{1, \dots, m\}$ be the subset of clusters such that $\ell \in L_s^\mu$ if and only if there exists no i^k such that $i^k \in \mu_s$ and $[i] = \ell$. Thus, the *effective capacity* (or *residual supply*) of s for the agents in clusters in L is $\min(|L_s^\mu \cap L|, q_s^\mu)$.

For illustration, suppose that a slot s has one unit of residual capacity and already contains a replica from cluster 1. Then $q_s^\mu = 1$ and cluster 1 is excluded from L_s^μ . If $L = \{1, 2\}$ and cluster 2 remains admissible at s , the slot contributes $\min(|L_s^\mu \cap L|, q_s^\mu) = 1$ unit of effective capacity for the replicas in clusters 1 and 2. Thus, residual capacity alone does not determine availability; the clusters already represented at the slot also matter.

Theorem 4 *A partial allocation μ admits a completion if and only if for each $L \subseteq \{1, \dots, m\}$,*

$$\sum_{\ell \in L} |A_\ell^\Gamma(\mu)| \leq \sum_{s \in S} \min(|L_s^\mu \cap L|, q_s^\mu).$$

The proof of Theorem 4 rests on a graph-theoretic reduction and two classical results: the Ford–Fulkerson max-flow/min-cut theorem and the integrality theorem for network flows. In Appendix A, we construct an auxiliary directed graph whose nodes and arc capacities encode the supply, demand, zone-coverage constraints, and admissible completions. Each feasible completion induces an integral flow whose value equals total residual demand, and every such integral flow yields a feasible completion by assigning the unassigned replicas of each cluster to the selected slots. Applying the Ford–Fulkerson theorem yields necessary and sufficient conditions for the existence of a feasible (possibly fractional) flow, and the integrality theorems guarantee that, under the capacity structure we impose, there exists a feasible flow that is integral. These two steps together produce the characterization stated in the theorem.

The inequalities in Theorem 4 need not be verified by enumerating all subsets of clusters. As established by the maximum-flow construction in its proof, a partial allocation admits a completion if and only if the corresponding residual network has a flow whose value equals total residual demand. Thus, completion feasibility can be tested by a single maximum-flow computation. The same completion

condition also has a matroidal interpretation as a matroid-union problem over slot-specific residual matroids; we discuss this connection in Appendix C.

Theorem 4 requires that for any subset of clusters, the number of unassigned replicas does not exceed the aggregate effective capacity of the slots.⁵ Consequently, a feasible allocation exists only if the empty allocation μ^0 satisfies:

$$\sum_{\ell \in L} |A_\ell^\Gamma(\mu^0)| \leq \sum_{s \in S} \min(|L_s^{\mu^0} \cap L|, q_s^{\mu^0}).$$

Note that at μ^0 , $A_\ell^\Gamma(\mu^0)$ is the set of all replicas in cluster ℓ , and $L_s^{\mu^0}$ is the set of all clusters. Moreover, for each s , $q_s^{\mu^0} = q_s$. These imply the following corollary of Theorem 4.

Corollary 1 *A feasible allocation exists if and only if for each $L \subseteq \{1, \dots, m\}$, we have*

$$\sum_{\ell \in L} \sum_{i \in A_\ell} d_i \leq \sum_{s \in S} \min(|L|, q_s).$$

The broader implication is that, under hard demand and capacity constraints, feasibility hinges critically on the specification of the zone-coverage relation—and thus on the structure of the induced clusters. In our pharmacy-scheduling application in Türkiye, these hard constraints bind because duty shifts are a major revenue source and the institution prioritizes each pharmacy’s centrally set demand. Corollary 1 therefore implies that ensuring an allocation satisfying both demand and capacity constraints requires careful design of the zone-coverage relation. In practice, however, the central authority fixes relatively dispersed subsets of pharmacies to meet hard demand and capacity constraints and reuses these subsets across shifts. The data, however, indicate that this heuristic occasionally fails the zone-coverage requirement (Section 4). Absent the formal treatment suggested by Corollary 1, many allocations that would be feasible under the same zone-coverage relation are effectively excluded, mirroring manual subset selection.

We maintain the condition in Corollary 1 as a standing assumption for the remainder of the analysis. Fixing a zone-coverage relation that (by assumption) guarantees feasibility, we assign agents to

⁵Observe that when for each s , $q_s = 2$ and $|L| = 1$, this inequality collapses to the one in Theorem 2.

slots dynamically via local decisions while preserving global feasibility at every step. The resulting mechanism satisfies replica-level no justified-envy and Pareto efficiency. It also satisfies replica-level pairwise stability when for each $s \in \mathcal{S}$, $q_s = 2$.⁶ To this end, we first define the set of available slots for i at μ :

$$S^\mu(i) = \{s \in \mathcal{S} : q_s^\mu > 0 \text{ and } [i] \notin \{[k] : k \in \mu_s\}\}.$$

To prevent a structural deadlock, we define the *feasibly available* slots for i at μ as follows:

$$S_{f-av}^\mu(i) = \{s \in S^\mu(i) : \forall L \subseteq \{1, \dots, m\}, \sum_{\ell \in L} |A_\ell^\Gamma(\mu \cup \{(i, s)\})| \leq \sum_{s' \in \mathcal{S}} \min(|L_{s'}^{\mu \cup \{(i, s)\}} \cap L|, q_{s'}^{\mu \cup \{(i, s)\}})\}.$$

By Theorem 4, a slot s belongs to $S_{f-av}^\mu(i)$ if and only if the tentative partial allocation $\nu = \mu \cup \{(i, s)\}$ admits a feasible completion. The displayed inequalities provide an analytical characterization of this condition. In the implementation of *GFSD*, however, membership in $S_{f-av}^\mu(i)$ is evaluated by the maximum-flow test described after Theorem 4, rather than by enumerating all subsets L . Equivalently, a locally available slot remains a candidate precisely when the tentative assignment of a replica to that slot leaves the residual problem completable.

That is, assigning i to a feasibly available slot ensures that for each subset of clusters L , the remaining demand does not exceed the remaining effective capacity. The mechanism processes the replicas sequentially according to the order $\succ^\Gamma$. We are now ready to define the mechanism we propose.

Generalized Feasible Serial Dictatorship (*GFSD*).

Step 0 Let μ^0 be the empty allocation.

Step k Consider the highest-priority replica under $\succ^\Gamma$ that has not yet been processed, and assign it to its most preferred slot in its feasibly available set. Denote the updated allocation by μ^k .

The mechanism *GFSD* terminates at Step $k = |\mathcal{A}^\Gamma|$, and its outcome is the allocation $\mu^{|\mathcal{A}^\Gamma|}$. For each $i \in \mathcal{A}$, $\mu(i) = \mu(i^1) \cup \dots \cup \mu(i^{d_i})$.

⁶In our running application, this approach replaces the central authority's heuristic in Türkiye with a principled, feasibility-guaranteeing method: rather than fixing specific subsets of pharmacies, we fix a zone-coverage relation and treat feasibility formally by selecting pharmacies for shifts systematically and dynamically.

Proposition 2 *On the cluster-approximated domain, GFSD can be implemented in time polynomial in the size of the original instance.*

With multi-unit demand, π_i orders individual slots. We extend it to bundles by listing each bundle's slots from best to worst under π_i and comparing bundles at the first rank where they differ, hence the *lexicographic extension*: For feasible allocations μ and μ' and each $i \in \mathcal{A}$, index the elements of $\mu(i)$ as $\mu_i^{(1)}, \dots, \mu_i^{(d_i)}$ such that $\mu_i^{(1)} \pi_i \mu_i^{(2)} \pi_i \dots \pi_i \mu_i^{(d_i)}$, and use the same best-to-worst indexing for $\mu'(i)$. Agent i *lexicographically prefers* $\mu(i)$ to $\mu'(i)$, denoted $\mu(i) \pi_i^{\text{lex}} \mu'(i)$, if there exists $t \in \{1, \dots, d_i\}$ such that:

$$\mu_i^{(t)} \pi_i \mu_i'^{(t)} \text{ and for each } r < t, \mu_i^{(r)} = \mu_i'^{(r)}.$$

The lexicographic extension is parsimonious in an informational sense. The primitive preference data consist only of each agent's ordinal ranking over individual slots. Alternative bundle models may introduce additional structure: additive representations require cardinal values or weights, responsiveness imposes restrictions on how slot rankings extend to bundles, and complementarity-based models specify interactions across slots. By contrast, the lexicographic extension orders each bundle from best to worst according to π_i and compares bundles at the first rank where they differ. It therefore introduces no cardinal utilities, weights, an additive-separability assumption, or parameters governing substitution or complementarity across slots. Thus, it is not the weakest possible extension in a normative sense, but a transparent complete ranking convention based only on the primitive slot ranking.

Since every replica has unit demand, the no justified-envy requirement in Definition 3 applies directly to the replica market. Specifically, we replace $\mathcal{A}$ and $\succ$ in that definition with $\mathcal{A}^\Gamma$ and $\succ^\Gamma$, respectively; each replica i^r inherits the slot preference π_i of its agent; and feasible alternatives are evaluated with respect to replica feasibility.⁷

We now define the standard efficiency criterion with respect to the lexicographic extension.

Definition 7 *An allocation μ is **Pareto efficient** if there exists no feasible allocation μ' such that for each $i \in \mathcal{A}$, $\mu'(i) \pi_i^{\text{lex}} \mu(i)$ or $\mu'(i) = \mu(i)$ and for at least one $j \in \mathcal{A}$ $\mu'(j) \pi_j^{\text{lex}} \mu(j)$.*

⁷No justified-envy can alternatively be formulated by comparing agents' complete bundles under the lexicographic extension. GFSD does not satisfy this bundle-level property in general; Appendix D provides the formal definition and a counterexample.

Throughout the multi-unit analysis, Pareto efficiency and incentive comparisons are evaluated under the lexicographic extension. No justified-envy is formulated at the replica level under the inherited slot preferences and the replica priority order $\succ^\Gamma$. For each $s \in \mathcal{S}$, pairwise stability is formulated at the replica level on the quota-two domain $q_s = 2$.

Remark 1 *The lexicographic extension respects the one-for-one improvements required by responsiveness and resolves bundle comparisons that those requirements alone leave undetermined.⁸ Pareto efficiency in Theorem 5 is evaluated under this complete lexicographic ranking.*

When for each $s \in \mathcal{S}$ $q_s = 2$, we also define pairwise stability analogously as before, replacing $\mathcal{F}$ with $\mathcal{F}^\Gamma$ and using the same pair-priority rule on replicas. Formally, a zone-covering replica pair (i^r, j^ℓ) blocks a feasible replica allocation μ at slot s if $s \pi_i^P \mu(i^r)$, $s \pi_j^P \mu(j^\ell)$, $(i^r, j^\ell) \in \mathcal{F}^\Gamma$, and $(i^r, j^\ell)_{\succ^\Gamma} \succ_c \mu_s$. A feasible replica allocation is pairwise stable at the replica level if no such pair exists. We are now ready to state our main result.

Theorem 5 *For each $\succ^\Gamma$ and each $(\pi_i)_{i \in \mathcal{A}}$, the outcome of GFSD satisfies feasibility, replica-level no justified-envy, and Pareto efficiency. If $q_s = 2$ for each $s \in \mathcal{S}$, it also satisfies replica-level pairwise stability.*

In contrast to the single-unit *FSD*, *GFSD* is not strategy-proof in general. The following example shows that the replica construction can create incentives for an agent to misreport.

Example 2 *Let $\mathcal{A} = \{i_1, i_2\}$ and $\mathcal{S} = \{s_1, s_2, s_3\}$. Slot capacities are $q_{s_1} = q_{s_2} = q_{s_3} = 1$. Agent demands are $d_{i_1} = 2$ and $d_{i_2} = 1$. Since each slot has capacity one, the zone-coverage constraint is irrelevant for this example. Let the priority ordering be $i_1 \succ i_2$. Then the set of replicas is $A^\Gamma = \{i_1^1, i_1^2, i_2^1\}$, and the replica order induced by $\succ^\Gamma$ is $i_1^1 \succ^\Gamma i_2^1 \succ^\Gamma i_1^2$.*

Let agent i_1 's true preference be $s_1 \pi_{i_1}^P s_2 \pi_{i_1}^P s_3$, and let agent i_2 's preference be $s_2 \pi_{i_2}^P s_3 \pi_{i_2}^P s_1$. Each replica inherits the preference of its agent. Under truthful reporting, i_1^1 chooses s_1 , then i_2^1 chooses s_2 , and finally i_1^2 receives s_3 . Thus, agent i_1 obtains the bundle $\{s_1, s_3\}$.

⁸Under the responsive extension, agent i prefers $\mu(i)$ to $\mu'(i)$ if the sets $\mu(i)$ and $\mu'(i)$ differ at only one slot, $s \in \mu(i)$ and $s' \in \mu'(i)$, respectively, such that $s \pi_i^P s'$.

Now suppose that agent i_1 instead reports $s_2 \hat{\pi}_{i_1}^P s_1 \hat{\pi}_{i_1}^P s_3$. Under this report, i_1^1 chooses s_2 , then i_2^1 chooses s_3 , and finally i_1^2 receives s_1 . Thus, agent i_1 obtains the bundle $\{s_1, s_2\}$.

Under agent i_1 's true preference and the lexicographic extension, $\{s_1, s_2\}$ is strictly preferred to $\{s_1, s_3\}$. Hence agent i_1 has a profitable misreport, and *GFSD* is not strategy-proof in general.

The example isolates the source of the incentive failure. Since all slots have capacity one, the zone-coverage constraint plays no role. Manipulability arises instead from the interaction between multi-unit demand and sequential allocation: an early replica of an agent can change the opportunity set faced by a later replica of the same agent. Thus, truth-telling should not be expected from *GFSD* in the unrestricted multi-unit domain. We therefore interpret *GFSD* as a completion-preserving allocation rule under reported or estimated preferences, rather than as a dominant-strategy implementation.

4 Pharmacy Duty-shift Scheduling: An Application

We evaluate the duty-shift scheduling problem in Section 1 using a data-driven simulation, comparing institutional practice with our mechanism on welfare and assignment structure. We use administrative records from Ankara, Türkiye (2015–2025): 3,905 shifts and 213 distinct pharmacies in the raw panel. After restricting attention to pharmacies with complete coordinate, priority, and cluster-recovery information, the simulation sample contains 198 pharmacies, with each record listing the shift, assigned pharmacies, and geolocations. All counterfactual mechanisms use the 198-pharmacy sample drawn from the Keçiören district of Ankara. Figure 4 depicts the full 198-pharmacy cluster-recovery sample. Accordingly, the reported tables and figures should not be interpreted as province-wide representations of Ankara. Shifts carry revenue coefficients (relative profitability reported by the Istanbul Chamber of Pharmacists, ICP), and capacities vary across the calendar, generating heterogeneous workloads across pharmacies.

Institutional practice does not formalize zone-coverage or priority; districts rely on informal partner sets and rotations that only partially reflect coverage. Coverage is implemented through shift-specific adjustments rather than a stable zone-coverage relation.⁹ While co-assignments are typically

⁹Across all co-assigned pharmacy pairs (pair-days), the median distance is 3.44 km and the 90th percentile is 5.54 km (2015–2025).

dispersed, Table 1 documents a number of ultra-proximate pairings in 2019–2025: 16 distinct pairs are assigned jointly at distances below 500 meters, appearing a total of 33 times; 75% of these occur only once, and the minimum observed distance is 51 meters.¹⁰ At the same time, one pair appears 10 times, consistent with the authority repeatedly reusing a small set of “known-feasible” exceptions to meet hard demand and capacity constraints. This motivates reconstructing zone-coverage from observed co-assignments and proximity before evaluating counterfactual mechanisms.

Table 1: Ultra-proximate co-assignments under the Status Quo schedule

Threshold	Unique pairs	Pair-days	Share once (%)	Max count	Min dist. (m)
< 500m	16	33	75.0	10	51

Notes: A pair-day is a shift-day in which two pharmacies are co-assigned. “Share once” is the fraction of unique pairs that appear exactly once in the panel. “Max count” is the largest number of appearances of any single pair below 500m. “Min dist.” is the minimum observed distance among co-assignments below 500m.

For counterfactual evaluation, we reconstruct primitives from data: infer zone-coverage by combining observed co-assignments with geographic proximity; generate preferences via a utility model using official duty-shift coefficients plus idiosyncratic noise; match multi-unit demand and shift capacities to their empirical values; and construct priorities from cumulative duty-shift coefficients (2015–2023). These steps yield a fully specified simulation environment that preserves the observed scheduling structure while enabling systematic counter-factuals. Reconstruction details are given in Appendix F.

4.1 Simulation Results

We evaluate four mechanisms in the reconstructed environment:

- the *Status Quo* (*SQ*), the manually constructed schedule by Ankara Chamber of Pharmacists,
- *GFSD*, our proposed mechanism that allows any cross cluster partnership
- *Fixed-Pair GFSD heuristic*, which applies historical partner restrictions locally while using the cluster-based completion test of *GFSD*, and

¹⁰These failures of the zone-coverage requirement underscore the importance of Corollary 1 and its explicit incorporation into the allocation mechanism.

- *Myopic Serial Dictatorship (MSD)*, a benchmark that respects local slot-capacity and zone-coverage constraints but does not test completion feasibility.

The Fixed-Pair *GFSD* heuristic checks the historical partner restriction against the current occupants of each candidate slot. Its maximum-flow test verifies completion under the cluster constraints but does not impose historical partner restrictions on future assignments. The procedure therefore does not guarantee completion under the combined constraints. Its outcomes are descriptive and do not isolate the inherent feasibility cost of the historical partner graph.

We summarize realized partnership patterns using total partners and core partners. These measures describe the breadth and repetition of co-assignments; they do not test the completion condition in Theorem 4 or identify the source of assignment failures. We compare these mechanisms on feasibility, welfare, and structural flexibility. The robustness grid combines eight cluster-recovery or geographic specifications with five preference calibrations, $\rho \in \{0.0, 0.3, 0.5, 0.7, 0.9\}$, yielding 40 scenarios. The eight specifications consist of five repeated co-duty thresholds, $t \in \{2, 3, 5, 7, 10\}$, and three alternative geographic criteria evaluated relative to the baseline recovery.

- *Feasibility.* *SQ* achieves full assignment via discretionary adjustments. The Fixed-Pair *GFSD* heuristic assigns, on average, 86.9% of replicas. Its unmatched replicas cannot be attributed solely to historical partner restrictions because its completion test does not incorporate those restrictions. Across 40 scenarios, *MSD* produces 1.78 unassigned replicas on average, whereas *GFSD* clears the market without failures. Among the counterfactual priority-based algorithms considered, *GFSD* is the only one that clears the market in every scenario.

The gap between *MSD* and *GFSD* varies across preference environments. *MSD* failures remain small on average, yet they persist in some scenarios because myopic assignments need not preserve feasible completion. The gap is the largest under more heterogeneous preference environments, thus under low correlation, preferences, where *MSD* is most prone to deadlocks, while *GFSD* clears the market for each preference profile. Further evidence is reported in Appendix E.

- *Welfare.* Pharmacy preferences are not elicited, so realized welfare under the status quo is not

Table 2: Average Mechanism Performance Across Robustness Scenarios

Mechanism	Yield (%)	Unassigned	Mean Rank	Median Rank	Top-3 (%)	Partners (≥ 1)	Core (≥ 2)
<i>Status Quo</i>	100.0	0.00	181.4	180.0	1.2	13.6	3.9
<i>Fixed-Pair GFSD</i>	86.9	206.70	100.9	61.9	9.1	11.0	6.7
<i>MSD</i>	99.9	1.78	29.0	26.1	15.7	28.0	2.1
<i>GFSD</i>	100.0	0.00	29.3	26.1	15.7	28.1	2.1

Notes: Entries report averages over the full robustness grid, consisting of alternative cluster reconstructions, alternative distance specifications, and alternative preference calibrations (40 scenarios in total). Welfare metrics for all mechanisms, including the status quo, are computed under the same simulated preference profiles; they should therefore be interpreted as model-implied counterfactual welfare measures. Observed welfare under the status quo is unavailable because preferences were not elicited. For mechanisms with unassigned replicas, rank-based welfare statistics are computed conditional on assigned replicas; unassigned counts are reported separately. Lower rank indicates better preference satisfaction. Total partners count all unique co-assignees (≥ 1), while core partners reflect persistent connections (≥ 2).

observed. The *SQ* allocation is observed, however, and we evaluate it under the same simulated preference profiles used for the counterfactual mechanisms. The reported rank-based statistics are therefore model-implied counterfactual welfare measures, not observed welfare outcomes.

Under this common preference scale, *SQ* performs poorly: its average rank is 181.4, its median rank is 180.0, and only 1.2% of assignments fall within the top three preferred duty shifts. Fixed-Pair *GFSD* also yields a poor performance (mean rank 100.9, median 61.9). The large gap between mean and median evidences a right-skewed distribution. These assigned-rank outcomes describe the locally partner-restricted heuristic and do not identify the separate effect of the historical partner restriction. Both *MSD* and *GFSD* generate superior outcomes, with nearly 16% of assignments falling in the top three preferred duty shifts. On average, their welfare levels are nearly identical on the assigned margin. *MSD* generates, on average, 1.78 unassigned replicas due to violations of the feasibility condition in Theorem 4. These failures arise in instances where demand concentrates on a small set of high-score duty shifts with limited capacity: early greedy assignments exhaust local slack and preclude feasible continuation. By contrast, *GFSD* explicitly guards against such feasibility-destroying moves through its look ahead feasibility check and therefore clears the market without failures.

- *Structural flexibility.* Table 2 reports substantial differences in realized partnership patterns. Relative

to *SQ*, *GFSD* produces more distinct partners and fewer core partners. The Fixed-Pair *GFSD* heuristic exhibits fewer distinct partners and more repeated partnerships. These statistics describe partner diversity and repetition, not global completion feasibility or a causal explanation of rank outcomes.

The averages in Table 2 mask variation across reconstructed environments, which we summarize in Appendix E.1. There, we report ranges and standard deviations across the robustness grid, including alternative repeated co-duty thresholds, distance specifications, and preference environments. The qualitative ranking of mechanisms remains unchanged throughout: *GFSD* clears the market in every tested scenario, whereas *MSD* leaves positive numbers of replicas unassigned in some scenarios and Fixed-Pair *GFSD* remains structurally constrained.

5 Further Applications

We present several applications in health services that fall squarely within our theoretical framework.

- *Home Health Care Workforce Scheduling.*

Home-health agencies assign nurses or carers to daily shifts that involve visiting patients at fixed addresses across metropolitan or rural regions. Feasibility depends not only on workload and travel time but on geographic-dispersion requirements: each shift must cover a subset of patient locations that collectively ensure service across predefined zones (e.g., north/south districts, urban/rural sectors) to meet response-time targets, equity goals, or contractual obligations. Caregivers express preferences over shifts, neighborhoods, patient types, and times—some preferring compact routes, others specific districts—so the assignment must jointly form geographically feasible bundles of visits and match them to workers in a way that respects both coverage constraints and ordinal preferences; this structure corresponds to shifts as admissible zone-covering subsets and sites (or workers) with ordinal preferences over assignments. (Mascolo, Martinez, and Espinouse, 2021)

- *Community and Public Health Outreach Services.*

Public health departments operate fixed clinics and outreach points across socio-economically distinct neighborhoods, and shift planning must ensure each service period includes sufficiently dispersed sites to guarantee equitable access—especially for vulnerable populations. Thus admissible shifts are governed by a geographic coverage structure: a feasible shift must represent multiple designated zones so no area is systematically underserved. Clinics and service teams also have preferences over days, times, travel distance, and community specialization, so scheduling becomes a matching problem with structured feasibility: only zone-covering combinations of sites form shifts, and among those feasible combinations, bilateral preferences determine the final allocation. (Jin, 2023)

- *Emergency Medical Services (EMS) and Ambulance Deployment.*

Emergency medical systems operate from a network of fixed stations across a city or region, and each shift must position ambulances and paramedic teams so coverage is geographically balanced and response times meet mandated limits. This creates explicit or implicit zone-coverage constraints: a shift is feasible only if the selected stations collectively cover all critical districts, so clustering units in one area can violate coverage elsewhere even when overall staffing is adequate. Paramedics express preferences over station assignments and shift types (day/night, central/peripheral), and stations may prefer particular team compositions or continuity; the allocation must therefore reconcile dispersion-based feasibility with bilateral preferences, making this a natural instance of a geographically constrained matching problem. (Li, Zhao, and Wyatt, 2011)

- *Multi-Facility Care Home or Hospital Staff Scheduling.*

In large urban systems with multiple long-term care homes or hospitals across city zones, staff scheduling must ensure each shift staffs a geographically dispersed subset of facilities—especially when floating or rotating personnel are used to stabilize service—so feasible assignments are governed by coverage rules that require minimum representation across zones to meet equity and regulatory standards. Nurses and other healthcare workers rank preferred shifts and locations (commuting, team familiarity, workload), and facilities express preferences (experienced staff, continuity of care),

making scheduling a preference-sensitive matching problem in which shifts are zone-covering bundles and both sides hold preferences over feasible assignments. (Anderson, Bodur, Rathwell, and Sarhangian, 2021)

6 Concluding Remarks

This paper studies scheduling problems in which cluster-approximated spatial distribution requirements and hard capacity constraints fundamentally shape the set of feasible allocations. In such environments, feasibility is not a local property that can be enforced slot by slot, but a global condition that depends on how assignments interact across clusters of agents and slots. As shown both theoretically and through application, this feature causes standard matching mechanisms—designed around local acceptance or greedy priority rules—to fail even when feasible allocations exist. To address this problem, we propose a priority-based mechanism (*GFSD*) that embeds a dynamic global feasibility check into the assignment process. The mechanism guarantees full assignment whenever feasibility is attainable while preserving priority respect, replica-level no justified-envy, and Pareto efficiency. It also satisfies replica-level pairwise stability when $q_s = 2$ for every $s \in S$. The no justified-envy guarantee protects individual replicas under the layered priority order $\succ^\Gamma$ and does not imply the corresponding property for original agents’ complete bundles under $\succ$.

Our empirical analysis of pharmacy duty-shift scheduling demonstrates the practical relevance of these results. Using administrative data to reconstruct the scheduling environment, we show that the locally partner-restricted heuristic and the myopic priority mechanism can generate assignment failures or poor model-implied rank outcomes under the simulated preference profiles. By enforcing generalized feasibility conditions, the proposed mechanism preserves full assignment in every reported scenario and improves model-implied rank outcomes without altering priorities or the maintained preference profiles. The simulations document differences in realized partnership patterns alongside assignment and rank outcomes, but do not isolate partnership structure as their cause. More broadly, our results establish a general design principle for constrained matching and scheduling problems in which feasible completion admits a tractable characterization.

Appendix A Proofs

Proof of Lemma 1. (Only if) Suppose $\mathcal{F}$ is symmetric and negatively transitive. Define a binary relation $\sim$ on $\mathcal{A}$ such that $i \sim j$ if and only if $i = j$ or $(i, j) \notin \mathcal{F}$. We show that $\sim$ is an equivalence relation. By definition, for each $i \in \mathcal{A}$, $i \sim i$ holds. Hence, $\sim$ is reflexive. Moreover, since $\mathcal{F}$ is symmetric, $(j, i) \in \mathcal{F}$ implies $(i, j) \in \mathcal{F}$. By definition of $\sim$, if $i \sim j$ and $i \neq j$, then $(i, j) \notin \mathcal{F}$. Thus, if $i \sim j$, then $j \sim i$. Hence, $\sim$ is symmetric. Lastly, if $i \sim j$ and $j \sim k$, and $i \neq j$, $j \neq k$, then $(i, j) \notin \mathcal{F}$ and $(j, k) \notin \mathcal{F}$. By negative transitivity of $\mathcal{F}$, $(i, k) \notin \mathcal{F}$. Thus, $i \sim k$, implying transitivity of $\sim$. Thus, $\sim$ is an equivalence relation. Let $\{A_1, \dots, A_m\}$ be its equivalence classes, that form a partition of $\mathcal{A}$, and $[i]$ denote the $\sim$ class of i . We claim that $(i, j) \in \mathcal{F}$ if and only if $[i] \neq [j]$. If $[i] = [j]$ and $i \neq j$, then $i \sim j$ implies $(i, j) \notin \mathcal{F}$; if $i = j$, the same conclusion follows from irreflexivity of $\mathcal{F}$. Conversely, if $[i] \neq [j]$, then $i \neq j$ and $i \not\sim j$, which implies $(i, j) \in \mathcal{F}$. Hence $\mathcal{F}$ is exactly the set of cross-class pairs.

(If) Suppose there exists a partition $\{A_1, \dots, A_m\}$ of $\mathcal{A}$ such that $(i, j) \in \mathcal{F}$ if and only if $[i] \neq [j]$. Symmetry of $\mathcal{F}$ follows from the definition. For negative transitivity, let $(i, j) \notin \mathcal{F}$ and $(j, k) \notin \mathcal{F}$. Then $[i] = [j]$ and $[j] = [k]$, hence $[i] = [k]$, which implies $(i, k) \notin \mathcal{F}$. ■

Proof of Lemma 2. Fix $A \subseteq \mathcal{A}$ and suppose $F_A \neq \emptyset$. Let i_A^* be the $\succ$ -maximal agent in the set $\{i \in A : \text{there exists } j \in A \text{ such that } (i, j)_\succ \in F_A\}$ and let j_A^* be the $\succ$ -maximal agent such that $(i_A^*, j_A^*)_\succ \in F_A$. Now take any $(i, j)_\succ \in F_A$. By the choice of i_A^* , we have $i_A^* \succeq i \succeq j$. If $(i_A^*, j) \in \mathcal{F}$, then the choice of j_A^* gives $j_A^* \succeq j$. Otherwise, j and i_A^* belong to the same cluster, whereas i belongs to a different cluster because $(i, j) \in \mathcal{F}$. Hence, $(i_A^*, i)_\succ \in F_A$, and $j_A^* \succeq i \succ j$. In either case, $i_A^* \succeq i$ and $j_A^* \succeq j$, so $(i_A^*, j_A^*)_\succ \succeq_c (i, j)_\succ$. Finally, if some $h \in A$ had higher priority than i_A^* , then h would form a zone-covering pair with at least one of i_A^* and j_A^* because those two agents belong to distinct clusters. The higher-priority member of that pair would be h , contradicting the definition of i_A^* . Thus, i_A^* is the $\succ$ -maximal agent in A . ■

Proof of Proposition 1. Let $\mathcal{F}$ be negatively transitive and $\{A_1, \dots, A_m\}$ be the induced clusters. For $A = \emptyset$, the choice is uniquely $\emptyset$. For nonempty A with $F_A = \emptyset$, i_A^* is unique. If $F_A \neq \emptyset$,

Lemma 2 implies F_A has a $\succeq_c$ – maximal pair $(i_A^*, j_A^*)_{\succ}$, hence $C_s(A)$ is uniquely determined. For substitutability, let $A \subseteq B$ and $i \in A$. Suppose $i \in C_s(B)$. We show $i \in C_s(A)$. If $i = i_B^*$, then it implies $i = i_A^*$. Thus, $i \in C_s(A)$. If $F_B \neq \emptyset$ and $i = j_B^*$, then there are two cases. First, suppose $i_B^* \in A$. In this case, i remains the highest-priority zone-covering partner to i_B^* in A . Thus, $i \in C_s(A)$. Second, suppose $i_B^* \notin A$. Since $i = j_B^*$ is the highest-priority agent in B whose cluster differs from $[i_B^*]$, every $k \in B$ with $k \succ i$ satisfies $[k] = [i_B^*]$. If i is the highest-priority agent in A , then $i = i_A^*$ and $i \in C_s(A)$. Otherwise, $i_A^* \succ i$, so $[i_A^*] = [i_B^*] \neq [i]$. Thus, $(i_A^*, i) \in \mathcal{F}$, and no agent in A with priority higher than i belongs to a cluster different from $[i_A^*]$. Therefore, $i = j_A^*$ and again $i \in C_s(A)$. ■

Proof of Theorem 1. Since every slot has quota two and must contain agents from distinct clusters, the existence of a feasible allocation implies $m \geq 2$. Suppose $m = 2$. Let A_1, A_2 be the clusters. Since a feasible allocation exists and each filled slot must be cross-class by Lemma 1, $|A_1| = |A_2| = |S|$. Suppose DA allocation is infeasible. Note that the priority-respecting choice function chooses two agents from a set only if there exists a zone-covering pair. Thus, infeasibility arises only if some slots are under-filled. Let i be the first agent in the priority ordering who remains unmatched. Agent i eventually proposes to every slot. At the end of DA , each slot is therefore either full or holds a single agent in $A_{[i]}$; otherwise, a slot with residual capacity and an incumbent from the other cluster would retain i as a compatible second agent. Hence, every slot contains an agent from $A_{[i]}$. Since each slot can accommodate at most one agent from a given cluster, this implies that all $|S|$ agents in $A_{[i]}$ are matched. Because $|A_{[i]}| = |S|$, no agent in $A_{[i]}$ can remain unmatched, contradicting that i is unmatched.

Suppose next that every cluster is a singleton. Then every pair of distinct agents is zone-covering, so the spatial constraint never binds and the priority-respecting choice function reduces to the standard quota-two priority choice rule. If an agent were unmatched at the end of DA , she would have been rejected by every slot, so every slot would hold two agents. The slots would then hold all $2|S| = |A|$ agents, contradicting the existence of an additional unmatched agent.

Suppose $m > 2$ and that at least one cluster is non-singleton. Let M be a largest cluster and choose distinct agents $i_1, i_2 \in M$. Since a feasible allocation exists, $|M| \leq |S|$. Moreover, every cluster other

than M has size at most $|S| - 1$: if another cluster also had size $|S|$, those two clusters would contain all $2|S|$ agents, contradicting $m > 2$. After removing i_1 and i_2 , there remain $2(|S| - 1)$ agents, with no cluster represented more than $|S| - 1$ times. These agents can therefore be partitioned into $|S| - 1$ zone-covering pairs. Indeed, a multiset of $2r$ cluster labels can be partitioned into r pairs of distinct labels whenever no label occurs more than r times; the result follows by repeatedly pairing agents from two currently largest clusters. Let these pairs be $B_1, \dots, B_{|S|-1}$, and let the slots be $s_0, s_1, \dots, s_{|S|-1}$. For each $k \geq 1$, both agents in B_k rank s_k first. Agents i_1 and i_2 rank s_0 first, and i_2 ranks the remaining slots after s_0 in any strict order. Complete all other preferences arbitrarily, and choose a common priority ordering in which i_2 has the lowest priority and $i_1 \succ i_2$. In the first round, each s_k , $k \geq 1$, holds the pair B_k , while s_0 holds i_1 and rejects i_2 because i_1 and i_2 belong to the same cluster. Agent i_2 is then rejected by every remaining slot, each of which already holds a zone-covering pair of higher-priority agents. DA terminates with i_2 unmatched and one vacant position at s_0 . Thus, although a feasible allocation exists, DA is not guaranteed to produce one. ■

Proof of Theorem 2. Suppose for each agent i , $d_i = 1$ and for each slot s , $q_s = 2$. We apply Theorem 4. In this case, $|A_\ell^\Gamma(\mu)| = |A_\ell(\mu)|$. For $L = \{\ell\}$, each vacant slot and each partially filled slot whose incumbent belongs to another cluster contributes one unit of effective capacity and all other slots contribute zero. Thus,

$$\sum_{s \in S} \min(|L_s^\mu \cap \{\ell\}|, q_s^\mu) = |S_{A_\ell}^\mu|.$$

For any L with $|L| \geq 2$, a vacant slot contributes two units of effective capacity, a partially filled slot contributes one, and a full slot contributes zero. The contribution of every slot is therefore its residual capacity. Aggregate balance gives

$$\sum_{s \in S} \min(|L_s^\mu \cap L|, q_s^\mu) = \sum_{s \in S} q_s^\mu = \sum_{\ell=1}^m |A_\ell(\mu)| \geq \sum_{\ell \in L} |A_\ell(\mu)|.$$

The inequality for $L = \emptyset$ is immediate. Hence, all nonsingleton inequalities in Theorem 4 hold automatically, and for each ℓ , its singleton inequalities are precisely $|A_\ell(\mu)| \leq |S_{A_\ell}^\mu|$. This proves both directions. ■

Proof of Theorem 3. Recall that $i_1 \succ i_2 \succ \dots \succ i_{|\mathcal{A}|}$, and for each ℓ , $D_\ell(\mu^0) \geq 0$ (an assumption such that a feasible allocation exists—see the footnote in the description of the *FSD* mechanism in Section 3.1.3). Let μ be the outcome of *FSD*.

Feasibility. By Theorem 2, it is enough to show that after each step k at *FSD*, for each ℓ , $D_\ell(\mu^k) \geq 0$. We argue this by induction. At $k = 0$, the inequality follows directly from the characterizing condition for the existence of a feasible allocation restated above, establishing the initial step of the induction.

Inductive step. For each ℓ , let $D_\ell(\mu^{k-1}) \geq 0$. Thus, μ^{k-1} admits a completion. Hence, i_k is assigned to some slot in at least one completion, implying $S_{f-av}^\mu(i_k) \neq \emptyset$. We assign i_k to her most preferred slot s in $S_{f-av}^\mu(i_k)$. We consider two cases of s being vacant or partially filled.

Suppose $s \in S_0^{\mu^{k-1}}$. Then, for cluster $A_{[i_k]}$, $|A_{[i_k]}(\mu^k)| = |A_{[i_k]}(\mu^{k-1})| - 1$ and $|S_{A_{[i_k]}}^{\mu^k}| = |S_{A_{[i_k]}}^{\mu^{k-1}}| - 1$. That is, there is one fewer unassigned agent, and slot s is no longer available to agents in $A_{[i_k]}$. Thus, $D_{[i_k]}(\mu^k) = D_{[i_k]}(\mu^{k-1}) \geq 0$. For cluster $\ell \neq [i_k]$, the assignment of i_k does not affect the set of unassigned agents in A_ℓ . While slot s moves from vacant to partially filled, the set $S_{A_\ell}^\mu$ of available slots remains unchanged; hence $D_\ell(\mu^k) = D_\ell(\mu^{k-1}) \geq 0$.

Suppose $s \in S_1^{\mu^{k-1}}(i_k)$. Then $(i_k, j_{\mu^{k-1}}(s)) \in \mathcal{F}$ and $|\mu_s^k| = 2$. For cluster $[i_k]$, $|A_{[i_k]}(\mu^k)| = |A_{[i_k]}(\mu^{k-1})| - 1$ and $|S_{A_{[i_k]}}^{\mu^k}| = |S_{A_{[i_k]}}^{\mu^{k-1}}| - 1$. Thus, $D_{[i_k]}(\mu^k) = D_{[i_k]}(\mu^{k-1}) \geq 0$. For cluster $[j_{\mu^{k-1}}(s)]$, neither the set of unassigned agents nor the set of available slots changes. For clusters $\ell \notin \{[i_k], [j_{\mu^{k-1}}(s)]\}$, the set of unassigned agents is unchanged, but the set of available slots is reduced by one. Thus, $D_\ell(\mu^k) = D_\ell(\mu^{k-1}) - 1$. Since $s \in S_{f-av}^{\mu^{k-1}}(i_k)$, the definition of the feasibly available set requires $D_\ell(\mu^{k-1}) > 0$ for every such cluster. Hence, $D_\ell(\mu^k) \geq 0$.

No justified-envy. Let μ be the outcome of *FSD*. Suppose μ does not satisfy no justified-envy. Then there exist $k' > k$ and a feasible allocation μ' such that $s = \mu(i_{k'}) \pi_{i_k}^P \mu(i_k)$, $\mu'(i_k) = s$, and for each $k'' < k$, $\mu'(i_{k''}) \pi_{i_{k''}}^R \mu(i_{k''})$. Thus, $s \notin S_{f-av}^{\mu^{k-1}}(i_k)$, otherwise the assignment $\mu(i_k)$ would not be a π_{i_k} -maximal feasibly available slot at Step k . Consider the partial allocation $\tilde{\mu}$ such that for each $k'' \leq k$, $\tilde{\mu}(i_{k''}) = \mu'(i_{k''})$, and for each $k'' > k$, $\tilde{\mu}(i_{k''}) = \emptyset$. Since μ' is feasible, $\tilde{\mu}$ admits a completion. Thus, for at least one $k'' < k$, $\mu(i_{k''}) \neq \mu'(i_{k''})$. Otherwise, $s \in S_{f-av}^{\mu^{k-1}}(i_k)$.

Let $l = \min\{l' : l' < k \text{ and } \mu'(i_{l'}) \neq \mu(i_{l'})\}$. Since $\mu'(i_l) \pi_{i_l}^R \mu(i_l)$ and the assignments differ,

$\mu'(i_l) \pi_{i_l}^P \mu(i_l)$. Thus, $\mu'(i_l) \notin S_{f-av}^{\mu^{l-1}}(i_l)$, otherwise the assignment $\mu(i_l)$ would not be a π_{i_l} -maximal feasibly available slot at Step l . By the choice of l , μ^{l-1} coincides with μ' on all higher-priority agents. Since μ' is feasible, assigning i_l to $\mu'(i_l)$ while retaining all earlier assignments admits μ' as a completion. By Theorem 2 and the definition of the feasibly available set, $\mu'(i_l) \in S_{f-av}^{\mu^{l-1}}(i_l)$, a contradiction. Thus, μ satisfies no justified-envy.

Pairwise stability. Suppose, toward a contradiction, that $(i, j)_{\succ}$ blocks the *FSD* outcome μ at slot s , where $i \succ j$. Because both blocking agents strictly prefer s to their assignments, neither belongs to μ_s . Pair-priority dominance therefore implies that i has higher priority than every agent in μ_s . Since *FSD* processes agents in priority order and never revokes an assignment, slot s was empty when i was processed. Hence $s \in S_0^{\mu^{k-1}}$. Since i strictly prefers s to $\mu(i)$, this contradicts the choice made by *FSD*.

Pareto efficiency. Suppose μ is not Pareto efficient. Then, there is a feasible μ' such that for each $i \in \mathcal{A}$, $\mu'(i) \pi_i^R \mu(i)$, and for at least one $j \in \mathcal{A}$, $\mu'(j) \pi_j^P \mu(j)$. Let k^* be the smallest index in the priority ordering such that $\mu(i^{k^*}) \neq \mu'(i^{k^*})$. Thus, for each $k < k^*$, $\mu(i^k) = \mu'(i^k)$. Since preferences over slots are strict, $\mu'(i^{k^*}) \pi_{i^{k^*}}^R \mu(i^{k^*})$ and $\mu'(i^{k^*}) \neq \mu(i^{k^*})$ imply $\mu'(i^{k^*}) \pi_{i^{k^*}}^P \mu(i^{k^*})$.

Since μ' is feasible and coincides with μ^{k^*-1} on all higher-priority agents, $\mu^{k^*-1} \cup \{(i^{k^*}, \mu'(i^{k^*}))\}$ admits a feasible completion. Thus, $\mu'(i^{k^*}) \in S_{f-av}^{\mu^{k^*-1}}(i^{k^*})$. But, *FSD* assigns i^{k^*} at Step k^* to a $\pi_{i^{k^*}}$ -maximal element of $S_{f-av}^{\mu^{k^*-1}}(i^{k^*})$. Since $\mu'(i^{k^*}) \pi_{i^{k^*}}^P \mu(i^{k^*})$, it leads a contradiction.

Strategy-proofness. Fix an agent i_k and fix the reports of all agents other than i_k . Let π_{i_k} be the true preference of i_k , and let $\widehat{\pi}_{i_k}$ be an arbitrary alternative report. Let μ^t denote the partial allocation after Step t when i_k reports truthfully, and let $\widehat{\mu}^t$ denote the partial allocation after Step t when i_k reports $\widehat{\pi}_{i_k}$, holding all other reports fixed.

We first show that the two runs coincide up to Step $k - 1$. At Step 0, both runs start from the empty allocation, so $\mu^0 = \widehat{\mu}^0$. Suppose that $\mu^{t-1} = \widehat{\mu}^{t-1}$ for some $t < k$. The agent processed at Step t is $i_t \neq i_k$, and this agent has the same reported preference in both runs. Moreover, the set of feasibly available slots $S_{f-av}^{\mu^{t-1}}(i_t)$ is determined only by the current partial allocation and by the primitive feasibility data, namely the cluster structure, capacities, and remaining unassigned agents.

It does not depend on the reported preference of i_k . Hence

$$S_{f-av}^{\mu^{t-1}}(i_t) = S_{f-av}^{\widehat{\mu}^{t-1}}(i_t).$$

Therefore, i_t chooses the same slot in both runs, and $\mu^t = \widehat{\mu}^t$. By induction, $\mu^{k-1} = \widehat{\mu}^{k-1}$.

It follows that when i_k is processed, the set of feasibly available slots is the same under truthful reporting and under the alternative report. Let this common set be

$$B_{i_k} = S_{f-av}^{\mu^{k-1}}(i_k) = S_{f-av}^{\widehat{\mu}^{k-1}}(i_k).$$

Under truthful reporting, FSD assigns i_k to the π_{i_k} -maximal element of B_{i_k} . Under the alternative report $\widehat{\pi}_{i_k}$, FSD assigns i_k to the $\widehat{\pi}_{i_k}$ -maximal element of the same set B_{i_k} . Since the truthful assignment is the best element of B_{i_k} according to i_k 's true preference, the assignment obtained under any alternative report is weakly worse according to π_{i_k} .

Finally, because this is a single-unit-demand environment, the assignment of i_k is fixed at Step k and cannot be changed by later steps of the mechanism. Thus, there exists no alternative report $\widehat{\pi}_{i_k}$ such that i_k obtains a strictly better assignment than under truthful reporting. By Definition 5, FSD is strategy-proof. ■

Proof of Theorem 4. (Only if) Let μ be a partial allocation admitting a completion μ' . Fix $L \subseteq \{1, \dots, m\}$ and s . At most q_s^μ replicas can be assigned to s due to capacity. Also, the zone-coverage allows at most one replica from each cluster in $L \cap L_s^\mu$. Thus, s can accommodate at most $\min(|L_s^\mu \cap L|, q_s^\mu)$ replicas from $\bigcup_{\ell \in L} A_\ell^\Gamma(\mu)$ at μ' . Summing over all slots yields:

$$\sum_{\ell \in L} |A_\ell^\Gamma(\mu)| \leq \sum_{s \in \mathcal{S}} \min(|L_s^\mu \cap L|, q_s^\mu).$$

(If) For each $L \subseteq \{1, \dots, m\}$, suppose $\sum_{\ell \in L} |A_\ell^\Gamma(\mu)| \leq \sum_{s \in \mathcal{S}} \min\{|L_s^\mu \cap L|, q_s^\mu\}$. To demonstrate that μ has a completion, we employ tools from graph theory. We construct a directed graph $G = (V, E)$ where V is the set of vertices and E is the set of edges. V consists of a source σ , a sink τ , for each

$\ell \in \{1, \dots, m\}$, a cluster vertex v_ℓ , and for each $s \in \mathcal{S}$, a slot vertex v^s :

$$V = \{\sigma, \tau\} \cup \{v_\ell : \ell \in \{1, \dots, m\}\} \cup \{v^s : s \in \mathcal{S}\},$$

and E consists of an edge from σ to each v_ℓ , an edge from v_ℓ to v^s if and only if cluster ℓ is admissible at s , and an edge from each v^s to τ :

$$E = \{(\sigma, v_\ell) : \ell \in \{1, \dots, m\}\} \cup \{(v_\ell, v^s) : \ell \in L_s^\mu \text{ and } s \in \mathcal{S}\} \cup \{(v^s, \tau) : s \in \mathcal{S}\}.$$

Let $k : E \rightarrow \mathbb{R}_+$ be the capacity function which associates each $e = (v, v') \in E$ of G a nonnegative real number $k(e)$, called the capacity of edge e . The capacities of edges of G are defined as $k(\sigma, v_\ell) = |A_\ell^\Gamma(\mu)|$, $k(v_\ell, v^s) = 1$, and $k(v^s, \tau) = q_s^\mu$. The network constructed in this way is depicted in Figure 1. A σ - τ flow is a function $f : E \rightarrow \mathbb{R}_+$ satisfying for each $e \in E$, $0 \leq f(e) \leq k(e)$, and flow

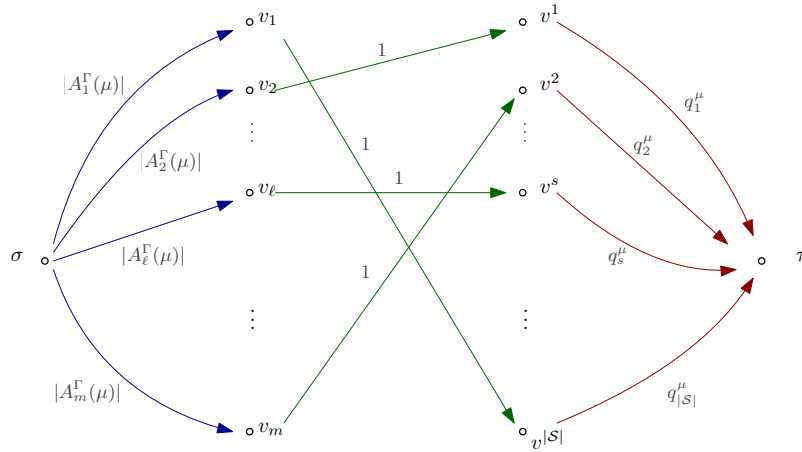


Figure 1

conservation such that for each $v \in V \setminus \{\sigma, \tau\}$: $\sum_{(v', v) \in E} f(v', v) = \sum_{(v, v'') \in E} f(v, v'')$. The *value* of a flow is the total flow sent out of σ . A partition of $V = U \cup \bar{U}$ is a (σ, τ) -cut if $\sigma \in U$ and $\tau \in \bar{U}$. Given a (σ, τ) -cut $(U, \bar{U})$, let $E' \subseteq E$ define the set of edges such that for each $e = (v, v') \in E'$, $v \in U$ and $v' \in \bar{U}$. We define the capacity of the cut as $k(U, \bar{U}) = \sum_{e \in E'} k(e)$. For a cut $(U, \bar{U})$, let

$S(U) = \{s : v^s \in U\}$, and $L = \{\ell : v_\ell \in U\}$. Note that the capacity of the cut $(U, \bar{U})$ is equal to

$$k(U, \bar{U}) = \sum_{\ell \notin L} |A_\ell^\Gamma(\mu)| + \sum_{s \in S \setminus S(U)} |L_s^\mu \cap L| + \sum_{s \in S(U)} q_s^\mu.$$

To prove our result, it suffices to demonstrate the existence of a flow of value $T = \sum_{\ell=1}^m |A_\ell^\Gamma(\mu)|$ in the graph we constructed. This is achieved using the well-known Ford-Fulkerson Theorem, which characterizes the value of the maximum flow.

Theorem 6 (*Ford-Fulkerson Theorem*) *Let $G = (V, A)$ be a directed graph. The maximum value of the σ - τ flow is equal to the minimum capacity among all (σ, τ) -cut.*

For each slot s , if $v^s \in U$ then the cut contributes q_s^μ , while if $v^s \in \bar{U}$ it contributes $|L_s^\mu \cap L|$; in both cases the contribution is at least $\min\{|L_s^\mu \cap L|, q_s^\mu\}$. That is,

$$\sum_{s \in S} \min\{|L_s^\mu \cap L|, q_s^\mu\} \leq \sum_{s \in S \setminus S(U)} |L_s^\mu \cap L| + \sum_{s \in S(U)} q_s^\mu. \quad (1)$$

Recall that we suppose for each L , $\sum_{\ell \in L} |A_\ell^\Gamma(\mu)| \leq \sum_{s \in S} \min\{|L_s^\mu \cap L|, q_s^\mu\}$. By combining this and (1) and adding $\sum_{\ell \notin L} |A_\ell^\Gamma(\mu)|$ to both sides, we obtain

$$\underbrace{\sum_{\ell \in L} |A_\ell^\Gamma(\mu)| + \sum_{\ell \notin L} |A_\ell^\Gamma(\mu)|}_T \leq \underbrace{\sum_{s \in S \setminus S(U)} |L_s^\mu \cap L| + \sum_{s \in S(U)} q_s^\mu + \sum_{\ell \notin L} |A_\ell^\Gamma(\mu)|}_{k(U, \bar{U})}$$

We now identify the minimum cut in Figure 2. Let $T = \sum_{\ell=1}^m |A_\ell^\Gamma(\mu)|$ denote the total residual demand for the set of all clusters. Thus, $k(U, \bar{U}) \geq T$. Since $(U, \bar{U})$ is arbitrary, the minimum capacity among all (σ, τ) -cut is greater than or equal to T . Consider the cut $(\{\sigma\}, V \setminus \{\sigma\})$. The only edges leaving σ are (σ, v_ℓ) such that:

$$k(\{\sigma\}, V \setminus \{\sigma\}) = \sum_{\ell \in \{1, \dots, m\}} k(\sigma, v_\ell) = \sum_{\ell \in \{1, \dots, m\}} |A_\ell^\Gamma(\mu)| = T.$$

Thus, the minimum cut's capacity is at most T . By the two bounds, the minimum cut capacity is

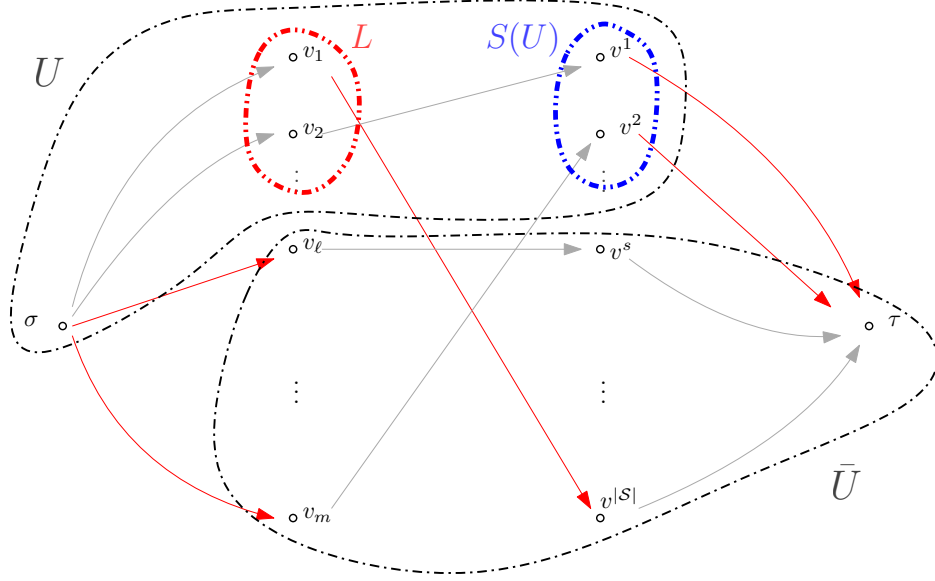


Figure 2

precisely T . By Theorem 6, the maximum value of $\sigma \rightarrow \tau$ flow is therefore T . To establish an integer flow, we invoke the following classical theorem from combinatorics (Schrijver, 2003).

Theorem 7 (Integer Theorem) *Let $G = (V, A)$ be a directed graph. If for each $a \in A$, $k(a)$ is integer, then there exists a maximum $\sigma \rightarrow \tau$ flow that is integral on every edge.*

By Theorem 7, there is an integer $\sigma \rightarrow \tau$ flow of value T . We interpret each unit of flow along the path $\sigma \rightarrow v_\ell \rightarrow v^s \rightarrow \tau$ as assigning one replica of cluster ℓ to s . The capacity constraints on edges (σ, v_ℓ) ensure that exactly $|A_\ell^\Gamma(\mu)|$ replicas of cluster ℓ are assigned in total, while the capacities on edges (v^s, τ) ensure that no slot s receives more than its residual capacity. Finally, the unit capacities on edges (v_ℓ, v^s) guarantee that no slot receives more than one replica from the same cluster. Thus, the flow induces a completion of μ . ■

Proof of Proposition 2. Let $R = |\mathcal{A}^\Gamma| = \sum_{i \in \mathcal{A}} d_i$ denote the total number of replicas. Since a feasible allocation exists, for each $i \in \mathcal{A}$, $d_i \leq |\mathcal{S}|$. Hence, $R = \sum_{i \in \mathcal{A}} d_i \leq |\mathcal{A}||\mathcal{S}|$, thus the expanded replica representation has size polynomial in the original instance. Feasibility also implies $q_s \leq m \leq |\mathcal{A}|$ for every $s \in \mathcal{S}$, so this conclusion continues to hold when demands and capacities are encoded in binary. The mechanism processes each replica exactly once and therefore terminates after R assignment steps. At each step, it tests at most $|\mathcal{S}|$ candidate slots.

By the residual-network construction in the proof of Theorem 4, each candidate-slot test can be carried out by a single maximum-flow computation. The corresponding network contains one source, one node for each of the m clusters, one node for each slot, and one sink. It therefore has $m + |\mathcal{S}| + 2$ vertices. It has m source-to-cluster arcs, at most $m|\mathcal{S}|$ cluster-to-slot arcs, and $|\mathcal{S}|$ slot-to-sink arcs. Hence, it has at most $m + m|\mathcal{S}| + |\mathcal{S}|$ arcs.

Thus, *GFSD* performs at most $R|\mathcal{S}|$ maximum-flow computations. Residual cluster demands, residual slot capacities, and cluster-presence indicators can be maintained throughout the mechanism, so each residual network can be constructed in $O(m|\mathcal{S}| + m + |\mathcal{S}|)$ time. Let $T_{\text{MF}}(V, E)$ denote the running time of a strongly polynomial maximum-flow algorithm on a network with V vertices and E arcs. The running time of *GFSD* is therefore

$$O(R|\mathcal{S}| [m|\mathcal{S}| + m + |\mathcal{S}| + T_{\text{MF}}(m + |\mathcal{S}| + 2, m|\mathcal{S}| + m + |\mathcal{S}|)]).$$

Since $R \leq |\mathcal{A}||\mathcal{S}|$ and $m \leq |\mathcal{A}|$, this bound is polynomial in the size of the original instance. Hence, *GFSD* is polynomial-time implementable on the cluster-approximated domain. ■

Proof of Theorem 5. Fix $\succ^\Gamma$. In the feasibility argument below, i^k denotes the replica processed at Step k . For each $k = 0, \dots, |\mathcal{A}^\Gamma|$, let μ^k denote the Step k output of *GFSD*, with the empty allocation μ^0 , and let $\mu = \mu^{|\mathcal{A}^\Gamma|}$ be the final outcome.

Feasibility. μ^0 satisfies the condition in Theorem 4, hence admits a completion. We prove by induction that, for each $k = 1, \dots, |\mathcal{A}^\Gamma|$, μ^k admits a completion.

Inductive step. Assume μ^{k-1} satisfies the condition in Theorem 4, and hence admits a completion $\tilde{\mu}$. Let $\tilde{s}$ denote the slot to which replica i^k is assigned under $\tilde{\mu}$. Since $\tilde{\mu}$ is feasible and is a completion of μ^{k-1} , assigning i^k to $\tilde{s}$ while retaining the assignments of μ^{k-1} respects both slot capacities and zone-coverage. Construct ν by assigning i^k to $\tilde{s}$ and retaining the assignments of μ^{k-1} without reassignment. Then $\tilde{\mu}$ is a feasible completion of ν . Hence, by Theorem 4 and the definition of the feasibly available set, $\tilde{s} \in S_{f-av}^{\mu^{k-1}}(i^k)$, so this set is nonempty. At Step k , *GFSD* chooses a slot $s \in S_{f-av}^{\mu^{k-1}}(i^k)$. By the definition of this set and Theorem 4, the updated allocation $\mu^k = \mu^{k-1} \cup \{(i^k, s)\}$ admits a

completion. Induction therefore establishes completion feasibility at each step. After all replicas have been processed, each replica is assigned exactly once and no assignments have been revoked. Since the final allocation still admits a completion and has no unassigned replicas, it is itself feasible.

No justified envy. Suppose toward a contradiction that the replica allocation μ produced by *GFSD* violates no justified-envy. Then, there exist replicas $(i^r, j^\ell)_{\succ \Gamma}$ with $\mu(j^\ell) \pi_i^P \mu(i^r)$ and a feasible replica allocation $\hat{\mu}$ with $\hat{\mu}(i^r) = \mu(j^\ell)$ such that, for each $h^q \succ^\Gamma i^r$, $\hat{\mu}(h^q) \pi_h^P \mu(h^q)$ or $\hat{\mu}(h^q) = \mu(h^q)$.

Let k^t be the $\succ^\Gamma$ -maximal replica in

$$\{i^r\} \cup \{h^q \in \mathcal{A}^\Gamma : h^q \succ^\Gamma i^r \text{ and } \hat{\mu}(h^q) \neq \mu(h^q)\}.$$

By the choice of k^t , each replica with priority higher than k^t is assigned identically under μ and $\hat{\mu}$. Moreover, $\hat{\mu}(k^t) \pi_k^P \mu(k^t)$. Indeed, if $k^t = i^r$, this follows from $\hat{\mu}(i^r) = \mu(j^\ell) \pi_i^P \mu(i^r)$. If $k^t \succ^\Gamma i^r$, it follows from the fact that $\hat{\mu}(k^t) \neq \mu(k^t)$ and that no replica with priority higher than i^r is made worse off under $\hat{\mu}$.

Consider the step of *GFSD* at which k^t is processed. Since all higher-priority replicas are assigned identically under μ and $\hat{\mu}$, the assignments made before this step coincide with the corresponding assignments under $\hat{\mu}$. Because $\hat{\mu}$ is feasible, assigning k^t to $\hat{\mu}(k^t)$ while retaining all earlier assignments admits $\hat{\mu}$ as a feasible completion. Hence, $\hat{\mu}(k^t)$ belongs to the feasibly available set of k^t at the step at which k^t is processed. Since $\hat{\mu}(k^t) \pi_k^P \mu(k^t)$, this contradicts the choice of a π_k -maximal feasibly available slot for k^t . Hence, μ satisfies replica-level no justified-envy.

Pairwise stability under quota two. Suppose that for each $s' \in \mathcal{S}$, $q_{s'} = 2$ and toward a contradiction, that the zone-covering replica pair $(i^r, j^\ell)_{\succ \Gamma}$ blocks the *GFSD* outcome μ at slot s . Pair-priority dominance implies that i^r has higher priority than both replicas in μ_s . It has at least the priority of the higher-priority incumbent and cannot itself be an incumbent since $s \pi_i^P \mu(i^r)$. Let μ^{k-1} be the partial allocation immediately before i^r is processed. Since *GFSD* never revokes an assignment, slot s is empty at μ^{k-1} . Thus, assigning i^r to s is locally admissible. Let $\nu = \mu^{k-1} \cup \{(i^r, s)\}$. By the feasibility argument above, μ^{k-1} satisfies every inequality in Theorem 4. We show that these inequalities are preserved at ν . If $[i] \in L$, assigning i^r to the empty slot s reduces residual demand in L by one

and reduces its effective capacity at s by one. The contributions from all other slots are unchanged. Thus, the inequality is preserved. If $[i] \notin L$ and $|L| \leq 1$, neither residual demand in L nor its effective capacity changes. Finally, suppose that $[i] \notin L$ and $|L| \geq 2$. Under quota two, a vacant slot contributes two units of effective capacity for L , a partially filled slot contributes one, and a full slot contributes zero. Thus, for each $s' \in \mathcal{S}$,

$$\min(|L_{s'}^{\mu^{k-1}} \cap L|, q_{s'}^{\mu^{k-1}}) = q_{s'}^{\mu^{k-1}}.$$

Hence, the effective capacity for L equals total residual capacity. Aggregate balance and the fact that the unassigned replica i^r lies outside L imply

$$\sum_{\ell' \in L} |A_{\ell'}^\Gamma(\mu^{k-1})| \leq \sum_{s' \in \mathcal{S}} q_{s'}^{\mu^{k-1}} - 1.$$

The tentative assignment leaves residual demand in L unchanged and reduces its effective capacity by exactly one, so this inequality is also preserved. Thus, ν admits a feasible completion by Theorem 4, and s is feasibly available when i^r is processed. Since $s \pi_i^P \mu(i^r)$, this contradicts the choice of a π_i -maximal feasibly available slot. Hence, no blocking pair exists.

Pareto efficiency. First, the replicas of each agent receive slots in best-to-worst order. Suppose otherwise that $r < t$ and $\mu(i^t) \pi_i^P \mu(i^r)$. Swapping their assignments in μ preserves feasibility because they belong to the same cluster. Since i^t is processed after i^r , the swapped allocation retains every assignment made before i^r is processed and provides a feasible completion assigning i^r to $\mu(i^t)$. This contradicts the choice of a π_i -maximal feasibly available slot for i^r . Hence, $\mu(i^r) = \mu_i^{(r)}$ for every i and r . Suppose the *GFSD* outcome μ is not Pareto efficient. Then, there is a feasible allocation μ' such that (i) for each $i \in \mathcal{A}$, either $\mu'(i) = \mu(i)$ or $\mu'(i) \pi_i^{\text{lex}} \mu(i)$, and (ii) there exists $j \in \mathcal{A}$ such that $\mu'(j) \pi_j^{\text{lex}} \mu(j)$. Represent μ' at the replica level by assigning each agent's r th-best slot to replica i^r , so that $\hat{\mu}(i^r) = \mu_i^{(r)'}$. This preserves feasibility. Let i^{*t} be the $\succ^\Gamma$ -maximal replica whose assignment differs between μ and $\hat{\mu}$. All higher-priority replicas have identical assignments. In particular, $\mu_{i^*}^{(r)'} = \mu_{i^*}^{(r)}$ for every $r < t$, whereas $\mu_{i^*}^{(t)'} \neq \mu_{i^*}^{(t)}$. Since $\mu'(i^*) \pi_{i^*}^{\text{lex}} \mu(i^*)$, this first difference implies $\mu_{i^*}^{(t)'} \pi_{i^*}^P \mu_{i^*}^{(t)}$. Consider the step k at which the replica i^{*t} is processed. By the choice of i^{*t} , μ^{k-1}

coincides with $\hat{\mu}$ on all assignments involving replicas with higher priority than i^{*t} in $\succ^\Gamma$. Since $\hat{\mu}$ is feasible, assigning i^{*t} to $\mu_{i^*}^{(t) \prime}$ while keeping all higher-priority assignments fixed admits a feasible completion. It implies $\mu_{i^*}^{(t) \prime} \in S_{f-av}^{\mu^{k-1}}(i^{*t})$. Since $\mu_{i^*}^{(t) \prime} \pi_{i^*}^P \mu_{i^*}^{(t)}$ and $\mu_{i^*}^{(t) \prime} \in S_{f-av}^{\mu^{k-1}}(i^{*t})$, the slot selected by the mechanism for replica i^{*t} cannot be the π_{i^*} -maximal element of $S_{f-av}^{\mu^{k-1}}(i^{*t})$. Thus, no such μ' exists, and μ is Pareto efficient. ■

Appendix B Capacity–Cluster Balance Beyond Quota Two

Theorem 1 identifies two distinct sources of DA’s uniform feasibility guarantee under quota two. The first is capacity–cluster balance: when the number of clusters equals the slot quota, every full feasible slot contains one agent from each cluster. The second is the nonbinding case in which every cluster is a singleton. This appendix extends the characterization to a homogeneous slot quota $q \geq 2$.

Consider a single-unit-demand environment with $K = |S|$ slots, each with quota q , so that $|A| = qK$. Let $A_1, \dots, A_m$ be the proximity clusters. We use the natural quota- q extension of the priority-respecting choice rule: a slot considers applicants in the common priority order and retains an agent if her cluster is not already represented, until either q agents have been retained or no additional cluster is represented. For each slot, this rule is precisely the fixed-priority greedy choice induced by the rank- q laminar matroid defined in Section 3.1.2. It is therefore substitutable for each $q \geq 2$ (Yokoi, 2019).

We use the following feasibility observation. If a population with cluster sizes $b_1, \dots, b_r$ satisfies $\sum_{\ell=1}^r b_\ell = qL$ and $b_\ell \leq L$ for every ℓ , then it can be assigned to L quota- q slots so that each slot contains agents from q distinct clusters. This follows from an integral-flow construction with capacity b_ℓ from the source to cluster ℓ , unit capacity from each cluster to each slot, and capacity q from each slot to the sink. The relevant Hall-type inequalities hold because, for every set R of clusters, $\sum_{\ell \in R} b_\ell \leq L \min\{|R|, q\}$.

Proposition 3 *Fix a single-unit-demand environment with homogeneous slot quota $q \geq 2$ and m proximity clusters.*

1. *If $m < q$, no complete feasible allocation exists.*

2. *Suppose that a complete feasible allocation exists. DA produces a complete feasible allocation for every common priority ordering and every preference profile if and only if either $m = q$ or every proximity cluster is a singleton.*
3. *Consequently, if the zone-coverage constraint is non-vacuous, DA has a uniform complete-feasibility guarantee if and only if $m = q$.*

Proof. If $m < q$, a slot can contain at most one agent from each of the m clusters and therefore cannot be filled to quota. Hence no complete feasible allocation exists.

Suppose that $m = q$ and that a complete feasible allocation exists. Every full feasible slot must contain one agent from each cluster. Since there are K slots, complete feasibility implies $|A_\ell| = K$ for every ℓ . Each slot therefore provides one cluster-specific position for every cluster. Under the priority-respecting choice rule, an acceptance from one cluster does not reduce the position available to another cluster. Restricted to any cluster, DA is the standard one-to-one deferred-acceptance procedure between the K agents in that cluster and the K slots. It assigns every agent, so every slot receives one agent from each of the q clusters and the outcome is complete and feasible.

Suppose instead that every cluster is a singleton. Then every pair of distinct agents is compatible, and DA reduces to the standard quota- q procedure. If an agent were unmatched, she would have been rejected by every slot, implying that every slot holds q agents. The slots would then hold all $qK = |A|$ agents, contradicting the existence of an additional unmatched agent. Hence DA produces a complete feasible allocation.

For the converse, suppose that $m > q$ and that at least one cluster is non-singleton. Let C be a largest cluster, with $|C| \geq 2$. Complete feasibility implies $|A_\ell| \leq K$ for every ℓ . Let H be the set of clusters other than C that contain exactly K agents. If $H \neq \emptyset$, maximality of C implies $|C| = K$. Hence $|H| \leq q - 2$; otherwise, C together with at least $q - 1$ other clusters of size K would exhaust the total population qK , leaving no agents for the additional clusters required by $m > q$. Because $m > q$ and $|H| \leq q - 2$, we can choose $q - 2$ distinct clusters other than C that include every cluster in H .

Select distinct agents $c, u \in C$ and one agent from each of the selected $q - 2$ clusters. After removing these q agents, the residual population has size $q(K - 1)$, and every residual cluster has size

at most $K - 1$. By the feasibility observation above, the residual agents can be assigned feasibly to $K - 1$ slots, each filled with q agents from distinct clusters.

Let s_0 be the remaining slot. Every residual agent ranks her assigned slot first. Agents c , u , and the selected agents from the other $q - 2$ clusters rank s_0 first. Complete all remaining preferences arbitrarily, with u ranking the other slots after s_0 , and choose a common priority ordering in which u has the lowest priority and $c \succ u$.

In the first round, each of the $K - 1$ residual slots receives and retains its designated feasible set of q agents. Slot s_0 receives q proposals from only $q - 1$ distinct clusters: c and u belong to C , while the other $q - 2$ applicants belong to distinct clusters. It retains c and the other $q - 2$ agents and rejects u . Agent u is subsequently rejected by every remaining slot, each of which is already filled with q higher-priority agents from distinct clusters. DA therefore terminates with u unmatched and one vacant position at s_0 , although a complete feasible allocation exists.

Thus DA has a uniform complete-feasibility guarantee exactly when $m = q$ or the spatial constraint is nonbinding because every cluster is a singleton. ■

Proposition 3 shows that the number two in Theorem 1 is not intrinsically special. With quota two, $m = 2$ is the balanced case, while three clusters are the first non-vacuous case in which the number of clusters exceeds the slot quota. When $m > q$ and the constraint is non-vacuous, a full slot can be formed from a strict subset of clusters, so locally admissible acceptances may alter residual cluster demands and eliminate all feasible completions. DA does not test for this residual effect; FSD and GFSD do so by requiring each tentative assignment to preserve completion feasibility.

Appendix C A Matroid-Union of Residual Completion: A Characterization Result

This appendix provides an alternative matroidal characterization of the residual-completion condition in Theorem 4. Although the feasibility constraint at each individual slot is laminar and induces a substitutable choice function, completing a partial allocation requires the unassigned replicas to be distributed jointly across the residual slot-specific matroids.

Fix a partial allocation μ . Let

$$U^\mu = \bigcup_{\ell=1}^m A_\ell^\Gamma(\mu)$$

denote the set of unassigned replicas. For each slot $s \in \mathcal{S}$, let $q_s^\mu = q_s - |\mu_s|$ denote its residual capacity, and let L_s^μ denote the set of clusters not represented at s under μ . Thus, replicas from cluster ℓ remain admissible at s if and only if $\ell \in L_s^\mu$.

For each slot $s \in \mathcal{S}$, define $M_s^\mu = (U^\mu, \mathcal{I}_s^\mu)$, where

$$\mathcal{I}_s^\mu = \left\{ I \subseteq U^\mu : \begin{array}{l} |I| \leq q_s^\mu, \\ |I \cap A_\ell^\Gamma(\mu)| \leq 1 \quad \text{for each } \ell \in L_s^\mu, \\ I \cap A_\ell^\Gamma(\mu) = \emptyset \quad \text{for each } \ell \notin L_s^\mu \end{array} \right\}.$$

Thus, $\mathcal{I}_s^\mu$ is the family of residual replica sets that can be added jointly to slot s : such a set cannot exceed the residual capacity of s , cannot contain two replicas from the same cluster, and cannot contain a replica from a cluster already represented at s .

Lemma 3 *For each partial allocation μ and slot $s \in \mathcal{S}$, the set system M_s^μ is a laminar matroid.*

Proof. Consider the family consisting of U^μ and the nonempty residual cluster sets $\{A_\ell^\Gamma(\mu) : A_\ell^\Gamma(\mu) \neq \emptyset\}$.

This family is laminar since the residual cluster sets are pairwise disjoint and each is contained in U^μ .

Assign capacity q_s^μ to U^μ . For each nonempty residual cluster set $A_\ell^\Gamma(\mu)$, assign capacity one when $\ell \in L_s^\mu$ and capacity zero when $\ell \notin L_s^\mu$. The subsets of U^μ satisfying these laminar capacity constraints are exactly the sets in $\mathcal{I}_s^\mu$. Hence, by the definition of a laminar matroid, M_s^μ is a laminar matroid. ■

A feasible completion of μ is equivalent to for each $s \in \mathcal{S}$, the existence of set $I_s \in \mathcal{I}_s^\mu$ such that $U^\mu = \bigcup_{s \in \mathcal{S}} I_s$. Indeed, I_s is the set of residual replicas assigned to slot s in the completion. Since total demand equals total slot capacity, $|U^\mu| = \sum_{s \in \mathcal{S}} q_s^\mu$. Moreover, for each $s \in \mathcal{S}$, $|I_s| \leq q_s^\mu$. Because the sets I_s partition U^μ , these inequalities jointly imply for each $s \in \mathcal{S}$, $|I_s| = q_s^\mu$. Thus, such a partition assigns each residual replica exactly once and fills every residual slot capacity while respecting the cluster restrictions.

To determine when such a partition exists, we apply Edmonds's matroid-union theorem, equivalently the matroid-partition theorem (Edmonds, 1968). Since this theorem is stated in terms of the rank functions of the slot-specific matroids, we first record the rank function of each M_s^μ . For each $X \subseteq U^\mu$, the rank function of M_s^μ is

$$r_s^\mu(X) = \min \{q_s^\mu, |\{\ell \in L_s^\mu : X \cap A_\ell^\Gamma(\mu) \neq \emptyset\}|\}. \quad (2)$$

Indeed, an independent subset of X can contain at most one replica from each cluster that is both represented in X and available at s , and it can contain at most q_s^μ replicas in total. Conversely, selecting one replica from each such cluster, up to q_s^μ replicas, attains the displayed bound.

Edmonds's matroid-union theorem now implies that a partition of U^μ into sets $I_s \in \mathcal{I}_s^\mu$, one for each $s \in \mathcal{S}$, exists if and only if for each $X \subseteq U^\mu$, $|X| \leq \sum_{s \in \mathcal{S}} r_s^\mu(X)$. The following proposition shows that these rank inequalities are exactly the cluster-based completion inequalities in Theorem 4.

Proposition 4 *For a fixed partial allocation μ , the following conditions are equivalent:*

(i) *For each $X \subseteq U^\mu$,*

$$|X| \leq \sum_{s \in \mathcal{S}} r_s^\mu(X).$$

(ii) *For each $L \subseteq \{1, \dots, m\}$,*

$$\sum_{\ell \in L} |A_\ell^\Gamma(\mu)| \leq \sum_{s \in \mathcal{S}} \min \{|L_s^\mu \cap L|, q_s^\mu\}.$$

Proof. Suppose first that condition (i) holds. Fix $L \subseteq \{1, \dots, m\}$, and let $L^+ = \{\ell \in L : A_\ell^\Gamma(\mu) \neq \emptyset\}$. Removing clusters with zero residual demand leaves the left-hand side of condition (ii) unchanged and weakly decreases its right-hand side. It is therefore sufficient to establish the inequality for L^+ .

Set $X = \bigcup_{\ell \in L^+} A_\ell^\Gamma(\mu)$. Because the residual cluster sets are pairwise disjoint, $|X| = \sum_{\ell \in L^+} |A_\ell^\Gamma(\mu)|$. Moreover, for each slot $s \in \mathcal{S}$, the clusters represented in X and available at s are exactly those in

$L_s^\mu \cap L^+$. Hence, by (2), $r_s^\mu(X) = \min \{|L_s^\mu \cap L^+|, q_s^\mu\}$. Applying condition (i) to X gives

$$\sum_{\ell \in L^+} |A_\ell^\Gamma(\mu)| \leq \sum_{s \in \mathcal{S}} \min \{|L_s^\mu \cap L^+|, q_s^\mu\}.$$

The same inequality therefore holds for L , so condition (ii) follows.

Conversely, suppose that condition (ii) holds. Fix $X \subseteq U^\mu$, and let $L(X) = \{\ell \in \{1, \dots, m\} : X \cap A_\ell^\Gamma(\mu) \neq \emptyset\}$. Since X contains only replicas from clusters in $L(X)$, $|X| \leq \sum_{\ell \in L(X)} |A_\ell^\Gamma(\mu)|$. Applying condition (ii) to $L(X)$ gives

$$|X| \leq \sum_{s \in \mathcal{S}} \min \{|L_s^\mu \cap L(X)|, q_s^\mu\}.$$

By the definition of $L(X)$ and (2), for each $s \in \mathcal{S}$, $\min \{|L_s^\mu \cap L(X)|, q_s^\mu\} = r_s^\mu(X)$. Consequently,

$$|X| \leq \sum_{s \in \mathcal{S}} r_s^\mu(X).$$

■

Relationship to Theorem 4. Condition (ii) in Proposition 4 is exactly the completion condition in Theorem 4. Together with Edmonds's theorem, the proposition therefore establishes that residual completion is a matroid-union problem: the unassigned replicas must be partitioned into sets that are independent in the respective residual slot matroids.

This characterization also clarifies the distinction between local substitutability and global completion. Each slot's fixed-priority greedy choice is substitutable because it operates under a laminar matroid. Nevertheless, a tentative assignment changes the common ground set of unassigned replicas and the associated residual slot matroids. At the receiving slot, it also reduces residual capacity and removes the assigned replica's cluster from the set of available clusters. Independently applied local choices therefore need not preserve the Edmonds rank inequalities required for a complete extension.

Relationship to the maximum-flow formulation. The matroid-union and maximum-flow formulations provide equivalent characterizations of the same residual-completion problem. The matroid-

union formulation places the problem within the established matroidal-constraint framework. The maximum-flow formulation exploits the additional structure of the cluster domain to provide a direct constructive and computational test: a single maximum-flow computation determines whether a partial allocation admits a completion, and an integral maximum flow directly identifies one. This residual maximum-flow test is the completion-feasibility oracle used at each step of *GFSD*.

Appendix D No Justified-Envy under Lexicographic Preferences

The no justified-envy property used in Theorem 5 is the direct application of Definition 3 to the unit-demand replica market. It evaluates priority claims according to the replica order $\succ^\Gamma$ and the slot preferences inherited by the replicas. An alternative is to formulate no justified-envy directly by comparing agents' complete bundles under the lexicographic extension.

Definition 8 *An allocation μ satisfies **strong no justified-envy** if for each $(i, j)_{\succ}$ with $d_i = d_j$ and $\mu(j) \pi_i^{\text{lex}} \mu(i)$ and each feasible allocation μ' with $\mu'(i) = \mu(j)$, there exists $k \succ i$ such that $\mu(k) \pi_k^{\text{lex}} \mu'(k)$.*

The *GFSD* outcome need not satisfy Definition 8. The failure arises because the layered replica order may place an early replica of a lower-priority agent ahead of a later replica of a higher-priority agent.

Example 3 *Let $\mathcal{A} = \{i, j, k\}$ and $\mathcal{S} = \{s_1, s_2, s_3\}$. Each agent demands two slots, and every slot has quota two. That is, $d_i = d_j = d_k = 2$ and $q_{s_1} = q_{s_2} = q_{s_3} = 2$.*

Let each agent form a singleton proximity cluster, so each pair of distinct agents is zone-covering. The agent priority order is $i \succ j \succ k$, and the agents' preferences are as follows: $s_1 \pi_i^P s_2 \pi_i^P s_3$, $s_2 \pi_j^P s_1 \pi_j^P s_3$, and $s_2 \pi_k^P s_1 \pi_k^P s_3$. The layered replica order is $i^1 \succ^\Gamma j^1 \succ^\Gamma k^1 \succ^\Gamma i^2 \succ^\Gamma j^2 \succ^\Gamma k^2$.

*Under *GFSD*, the replicas are assigned as follows:*

$$\mu(i^1) = s_1, \quad \mu(j^1) = s_2, \quad \mu(k^1) = s_2,$$

$$\mu(i^2) = s_3, \quad \mu(j^2) = s_1, \quad \mu(k^2) = s_3.$$

At every step, the indicated assignment is the most-preferred feasibly available slot; the remaining assignments displayed above provide a feasible completion of the corresponding partial allocation. The induced allocation is therefore

$$\mu(i) = \{s_1, s_3\}, \quad \mu(j) = \{s_1, s_2\}, \quad \mu(k) = \{s_2, s_3\}.$$

Under agent i 's lexicographic preference, $\mu(j) \pi_i^{\text{lex}} \mu(i)$, since the two bundles have the same best slot s_1 , while i prefers s_2 to s_3 . Consider the alternative allocation

$$\widehat{\mu}(i) = \{s_1, s_2\}, \quad \widehat{\mu}(j) = \{s_1, s_3\}, \quad \widehat{\mu}(k) = \{s_2, s_3\}.$$

This allocation is feasible: every agent receives two slots, every slot receives two agents, and every slot contains agents from distinct proximity clusters. Moreover, $\widehat{\mu}(i) = \mu(j)$. Since i is the highest-priority agent, there exists no agent $h \succ i$ such that $\mu(h) \pi_h^{\text{lex}} \widehat{\mu}(h)$. Hence, the *GFSD* outcome violates strong no justified-envy.

Example 3 does not contradict the fairness result in Theorem 5. The slot s_2 desired by i at the second position of her bundle is occupied by j^1 and k^1 , both of whom precede i^2 under $\succ^\Gamma$ and rank s_2 first. In particular, the alternative allocation $\widehat{\mu}$ moves j^1 away from s_2 and therefore does not preserve the assignments of all replicas preceding i^2 . The example thus separates priority protection for marginal units under $\succ^\Gamma$ from priority protection for complete bundles under $\succ$.

Appendix E The value of *GFSD*: Further Evidence

We also examine how preference correlation interacts with congestion (Table 3). For each ρ , preferences are generated by combining official duty shift coefficients with idiosyncratic noise (explained in detail in Section F.2 below). *Peak pressure*—the maximum ratio of top-three replica demand to duty shift capacity—describes preference concentration relative to slot capacity. Across all calibrations, *MSD* occasionally produces assignment failures, which vary across preference environments. It reflects violations of the completion condition in Theorem 4. In contrast, *GFSD* always clears the

market, demonstrating robustness to demand concentration.

The Fixed-Pair *GFSD* heuristic leaves roughly 205–209 replicas unassigned on average across the preference calibrations. These outcomes reflect a procedure that checks historical partner restrictions locally but evaluates future completion only under the cluster constraints. They therefore do not establish whether the combined constraints admit a complete allocation or isolate the inherent cost of restricting partnerships to historically observed pairs.

Table 3: Preference Alignment, Congestion, and Assignment Failures

Preference	Mean	Peak Pressure	<i>MSD</i>	<i>GFSD</i>	Fixed-Pair <i>GFSD</i>
Scenario	Spearman	(Top-3)	Unassigned	Unassigned	Unassigned
$\rho = 0.0$	0.000	18.0	3.00	0.00	205.38
$\rho = 0.3$	0.237	42.0	1.62	0.00	207.38
$\rho = 0.5$	0.379	98.0	1.75	0.00	208.50
$\rho = 0.7$	0.514	170.8	1.12	0.00	206.38
$\rho = 0.9$	0.668	282.8	1.38	0.00	205.88

Notes: ρ is the loading parameter in the preference-generating model. Mean Spearman reports the average realized rank correlation between pharmacy-specific slot rankings and the common coefficient ranking in the 198-pharmacy simulation sample, rounded to three decimals; tied coefficients receive average ranks. Peak pressure is the maximum, across duty shifts, of aggregate top-three replica demand divided by slot capacity. It is computed once per preference profile and is invariant across cluster-recovery and distance specifications. Unassigned counts are averaged across the eight cluster-recovery and distance specifications for each preference profile.

Figure 3 reports the model-implied assigned-rank distributions for the baseline simulation.

E.1 Robustness and Simulation Variability

Table 4 summarizes the robustness exercises that vary the recovered cluster structure and the distance specification used in the recovery stage. The qualitative ranking of mechanisms is unchanged. *GFSD* clears the market in every specification. *MSD* remains close to *GFSD* in many cases, but it lacks a uniform clearing guarantee and leaves up to four replicas unassigned under alternative cluster reconstructions. The Fixed-Pair *GFSD* heuristic leaves roughly 195–219 replicas unassigned across these

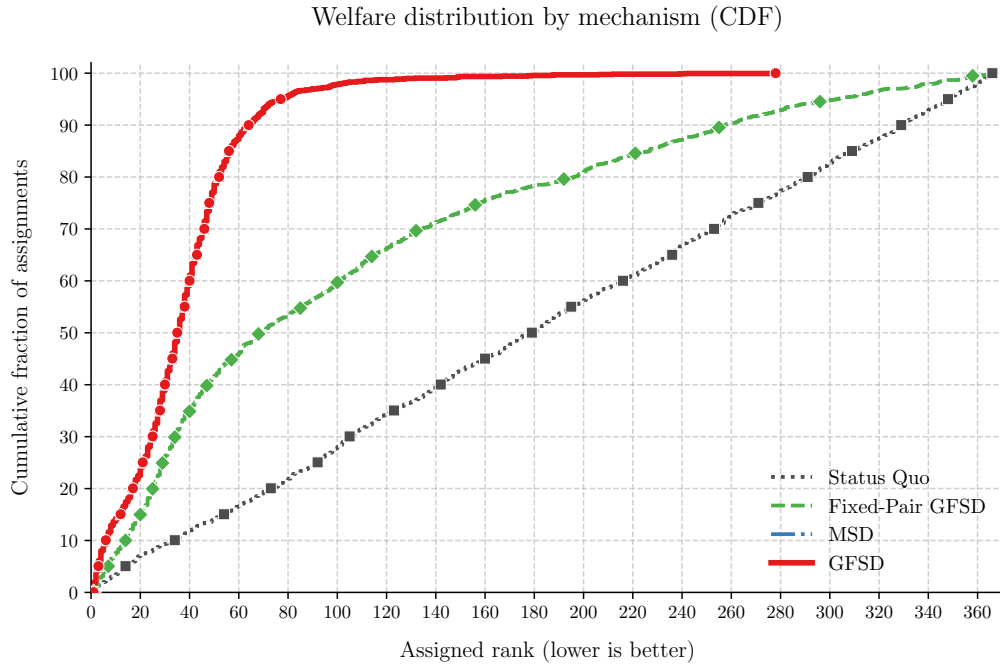


Figure 3: Model-Implied Assigned-Rank Distributions under Simulated Preferences.

The figure reports cumulative distributions of assigned ranks under *SQ*, Fixed-Pair *GFSD*, *MSD*, and *GFSD* for the baseline cluster reconstruction ($t = 3$) and preference calibration ($\rho = 0.7$). Lower ranks indicate better preference satisfaction. Rank distributions for mechanisms with unassigned replicas are conditional on assigned replicas.

robustness exercises.

Table 4: Robustness across Cluster and Distance Specifications

Robustness dimension	Specifications	<i>GFSD</i>	<i>MSD</i>	Fixed-Pair <i>GFSD</i>
		Unassigned	Unassigned	Unassigned
Cluster reconstruction	Repeated co-duty thresholds $t \in \{2, 3, 5, 7, 10\}$, with recovered $K^* \in \{10, 11, 12\}$	0–0	0–4	195–216
Distance specification	Linear, unpenalized squared, and Haver- sine distance, hold- ing $t = 3$, $K^* = 11$, and $\rho = 0.7$ fixed	0–0	0–2	196–219

Notes: Both robustness dimensions hold the preference calibration fixed at $\rho = 0.7$. Ranges summarize variation across the specifications listed in each row; Table 5 reports variation across the full 40-scenario grid.

Because pharmacy preferences are not elicited in the administrative data, the simulations are counterfactual exercises under reconstructed environments rather than estimates based on observed preference data. We therefore summarize sensitivity to the maintained reconstruction and calibration choices. Table 5 reports scenario-level standard deviations and ranges for the principal feasibility and rank-based welfare outcomes across the 40-scenario grid. These statistics describe variation across the specifications considered; they are not confidence intervals or sampling standard errors.

Table 5: Variation across Robustness Scenarios

Mechanism	Unassigned			Mean Rank			Top-3 (%)		
	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
<i>Status Quo</i>	0.00	0.00	0–0	181.4	2.9	179.3–186.9	1.2	0.2	0.8–1.3
<i>Fixed-Pair GFSD</i>	206.70	7.18	194–220	100.9	12.5	86.7–124.3	9.1	7.7	0.9–21.4
<i>MSD</i>	1.78	1.89	0–8	29.0	20.5	8.9–64.8	15.7	11.4	1.9–32.9
<i>GFSD</i>	0.00	0.00	0–0	29.3	20.4	9.2–65.4	15.7	11.4	1.9–32.8

Notes: Entries summarize variation across the full robustness grid of 40 scenarios. Ranges report the minimum and maximum scenario-level values. Rank-based welfare statistics are computed conditional on assigned replicas; unassigned counts are reported separately. Lower rank indicates better preference satisfaction.

The qualitative comparison of mechanisms is stable across the grid. *GFSD* clears the market in every scenario, with zero variation in unassigned replicas. *MSD* delivers nearly identical assigned-rank outcomes, but leaves up to eight replicas unassigned in some scenarios. The Fixed-Pair *GFSD* heuristic leaves between 194 and 220 replicas unassigned across scenarios.

Appendix F Data Generation

F.1 Zone-Coverage Relation

While geographical coordinates are available, Euclidean distance alone fails to induce a negatively transitive binary relation, consequently, a partition. Data reveals only a subset of zone-covering partners that are frequently matched together. This leaves the relation between the unobserved partners ambiguous.

We adopt a constructive approach to recover the zone-coverage. For a repeated co-duty threshold t , let E_t denote the graph whose edges are pharmacy pairs that were historically assigned to the same duty shift at least t times. Each edge $(i, j) \in E_t$ is treated as hard evidence that pharmacies i and j belong to different service zones. Equivalently, for each $(i, j) \in E_t$, the recovered partition C_t must satisfy $C_t(i) \neq C_t(j)$. Among partitions satisfying these hard cross-zone constraints and the cluster-

size bounds, we select geographically compact partitions.

The recovery proceeds in two steps. First, for each threshold t , we compute the smallest certified feasible number of clusters $K_t^{\min}$ satisfying the hard co-duty constraints and the size bounds. Second, beginning from $K_t^{\min}$, we solve a compactness grid over larger values of K and select the smallest K_t^* after which geographic coherence stabilizes. We repeat the recovery for $t \in \{2, 5, 7, 10\}$. These alternatives are visualized in Figure 4.

We also assess sensitivity to the geographic criterion used in the recovery procedure. Holding the co-duty threshold and the number of clusters fixed at their baseline values, $t = 3$ and $K^* = 11$, we compare the baseline squared-distance specification with a far-distance penalty against three alternatives: a linear-distance specification, a squared-distance specification without the far-distance penalty, and a squared Haversine-distance specification with the far-distance penalty. These exercises isolate sensitivity to the functional form of the geographic compactness criterion while holding the co-duty constraints and the number of recovered clusters fixed.

The induced relation $\mathcal{F}$ contains exactly the distinct cross-cluster pairs and is therefore symmetric, irreflexive, and negatively transitive. Indeed, the complement of $\mathcal{F}$ is the equivalence relation of belonging to the same recovered service zone. Hence, the recovered relation is both theoretically consistent with the partition structure required by the mechanism and empirically grounded in the district's geography and observed co-duty data.

Figure 4 provides an illustrative visualization of the recovered service zones in the Ankara/Keçiören district-level subsample under alternative repeated co-duty thresholds. The clusters exhibit clear geographical coherence across specifications. Panel B corresponds to the baseline partition, $t = 3$ and $K^* = 11$, used in the main mechanism simulations. The remaining panels show the robustness partitions for $t \in \{2, 5, 7, 10\}$. Colored regions correspond to convex hulls of the recovered clusters and are included only as visual aids; they are not imposed constraints in the recovery procedure. Pharmacies within the same recovered service zone are geographically concentrated, and the broad spatial structure of the partition remains stable as the hard co-duty threshold varies.

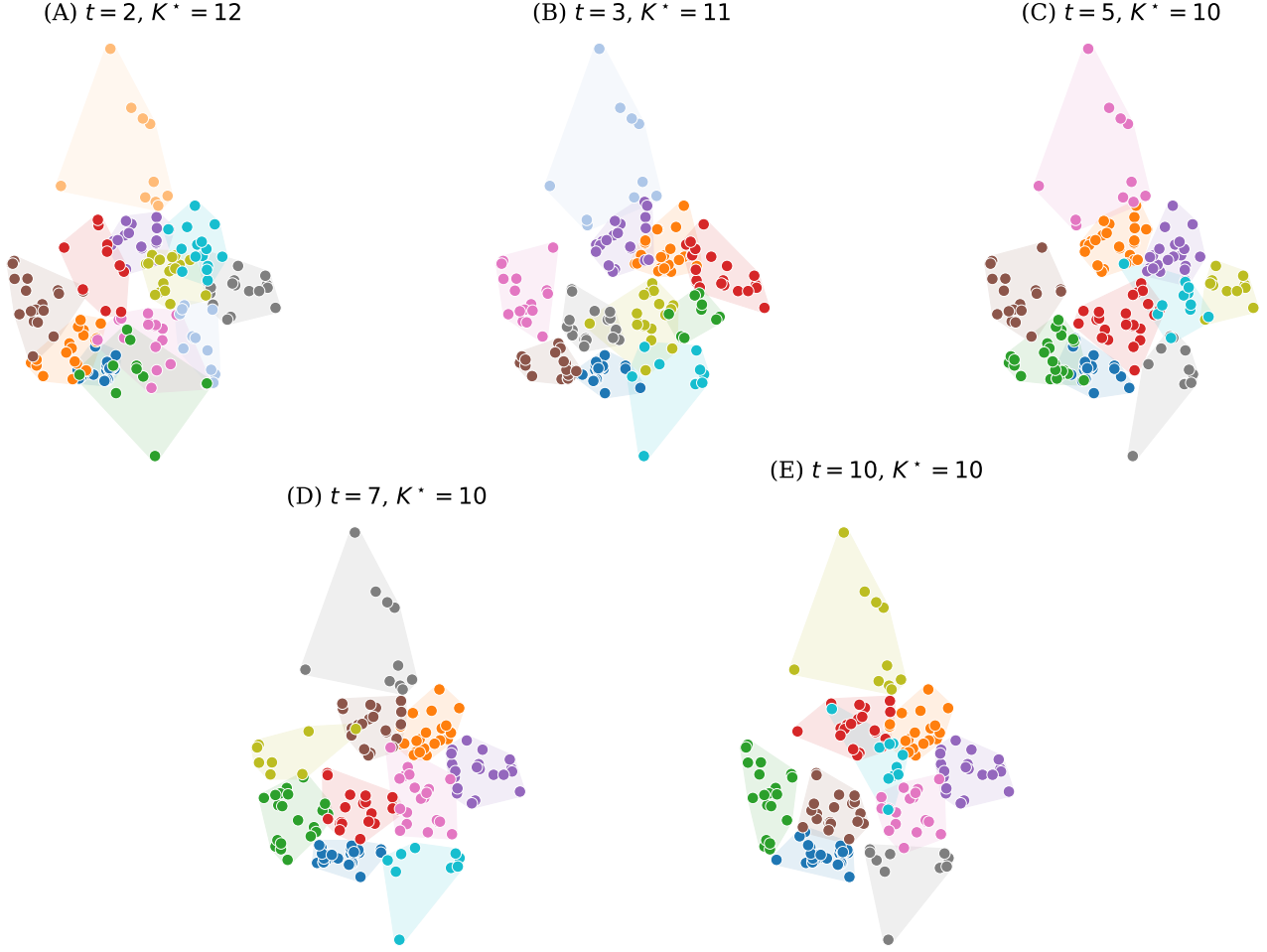


Figure 4: **Recovered service zones.**

The figure depicts the full 198-pharmacy cluster-recovery sample from the Ankara/Keçiören district-level subsample.

F.2 Preferences and Priority Ordering

We construct a Data-Generating Process (DGP) that produces synthetic but behaviorally disciplined preference profiles. The financial attractiveness of the slot and an idiosyncratic component derives the utility of a pharmacy for a specific slot. Formally, the utility of pharmacy i for a slot s , denoted as $u_{i,s}$, is as follows:

$$u_{i,s} = \rho \cdot Z_s + \sqrt{1 - \rho^2} \cdot \epsilon_{i,s}$$

Z_s represents the financial attractiveness of slot s , common to all pharmacies. We proxy Z_s using the slot coefficients provided by a Region Representative of the Istanbul Chamber of Pharmacists.

The coefficients summarize expected patient flow and workload, and thus correlate with the revenue potential of a slot. We normalize Z_s to have mean zero and unit variance.

We capture unobserved pharmacy-specific shocks that affect desirability of a slot by $\epsilon_{i,s} \sim \mathcal{N}(0, 1)$. These shocks are assumed independent identically distributed across i and s . The scalar $\rho \in [0, 1]$ adjusts the extent of preference alignment in the market. When ρ is high, preferences are largely determined by the revenue gradient embedded in Z_s . When ρ is low, pharmacies' preferences are driven mostly by idiosyncratic factors. The ordinal preference relation π_i for each pharmacy is derived by ranking the slots according to utilities such that for each $i \in \mathcal{A}$ and $s, s' \in \mathcal{S}$, $s \pi_i s'$ if and only if $u_{i,s} > u_{i,s'}$.

For each loading parameter ρ , we generated $N = 5000$ candidate preference profiles in the original 210-pharmacy calibration universe. For each candidate, we computed the average Spearman rank correlation between pharmacy-specific slot rankings and the common coefficient ranking, and selected the realization whose average correlation was closest to ρ . The selected profiles were then restricted to the 198 pharmacies used in the mechanism simulations. The loading parameter ρ is distinct from the realized rank correlation. Table 3 reports the realized average Spearman correlations computed over this 198-pharmacy sample.

The priority ordering $\succ$ is determined by slot coefficients. For each pharmacy, we compute the cumulative coefficient over 2015–2023 and rank pharmacies from lowest to highest. This ranking serves as the empirical analogue of the exogenous priority order in the theoretical model and is consistent with institutional rotation: pharmacies historically assigned to lower-valued slots receive strictly higher priority.

References

- AFACAN, M. O., A. DUR, AND U. DUR (2025): “Seat Allocation Problem in Public Transportation,” *Scandinavian Journal of Economics*, forthcoming.
- ANDERSON, M., M. BODUR, S. RATHWELL, AND V. SARHANGIAN (2021): “Optimization Helps Scheduling Nursing Staff at the Long-Term Care Homes of the City of Toronto,” working paper.

- BARMAN, S., AND A. BISWAS (2020): “Fair Division Under Cardinality Constraints,” working paper.
- ECHENIQUE, F., AND J. OVIEDO (2006): “A theory of stability in many-to-many matching,” *Theoretical Economics*, 1, 233–273.
- ECHENIQUE, F., AND M. B. YENMEZ (2015): “How to Control Controlled School Choice,” *American Economic Review*, 105(8), 2679–94.
- EDMONDS, J. (1968): “Matroid partition,” in *Mathematics of the Decision Sciences, Part I*, ed. by G. B. Dantzig, and J. Veinott, Arthur F., vol. 11 of *Lectures in Applied Mathematics*, pp. 335–345. American Mathematical Society, Providence, RI.
- FIFE, T., AND J. OXLEY (2017): “Laminar matroids,” *European Journal of Combinatorics*, 62, 206–216.
- FRAGIADAKIS, D., AND P. TROYAN (2017): “Improving matching under hard distributional constraints,” *Theoretical Economics*, 12(2), 863–908.
- HATFIELD, J., AND F. KOJIMA (2010): “Substitutes and Stability for Matching with Contracts,” *Journal of Economic Theory*, 145, 1704–1723.
- HATFIELD, J., AND P. MILGROM (2005): “Matching with Contracts,” *American Economic Review*, 95, 913–935.
- IMAMURA, K., AND Y. KAWASE (2024): “Efficient matching under general constraints,” *Games and Economic Behavior*, 145(C), 197–207.
- (2025): “Efficient and strategy-proof mechanism under general constraints,” *Theoretical Economics*, 20, 481–509.
- JIN, Z. (2023): “Home Health Care Problem Related to Routing and Scheduling,” *Frontiers in Business, Economics and Management*, 9 (2), 240–242.

- KAMADA, Y., AND F. KOJIMA (2015): “Efficient Matching under Distributional Constraints: Theory and Applications,” *American Economic Review*, 105(1), 67–99.
- (2018): “Stability and strategy-proofness for matching with constraints: A necessary and sufficient condition,” *Theoretical Economics*, 13(2), 761–793.
- (2024): “Fair Matching under Constraints: Theory and Applications,” *The Review of Economic Studies*, 91(2), 1162–1199.
- KOJIMA, F., N. SUN, AND N. N. YU (2020): “Job Matching under Constraints,” *American Economic Review*, 110(9), 2935–47.
- KOJIMA, F., A. TAMURA, AND M. YOKOO (2018): “Designing matching mechanisms under constraints: An approach from discrete convex analysis,” *Journal of Economic Theory*, 176, 803–833.
- LI, X., Z. ZHAO, AND T. WYATT (2011): “Covering models and optimization techniques for emergency response facility location and planning: a review,” *Mathematical Methods of Operations Research*, 74, 281–310.
- MANEA, M. (2007): “Serial dictatorship and Pareto optimality,” *Games and Economic Behavior*, 61, 316–330.
- MASCOLO, M. D., C. MARTINEZ, AND M.-L. ESPINOUSE (2021): “Routing and scheduling in Home Health Care: A literature survey and bibliometric analysis,” *Computers & Industrial Engineering*, 158.
- SCHRIJVER, A. (2003): *Combinatorial Optimization: Polyhedra and Efficiency*. Springer-Verlag.
- SÖNMEZ, T. (2025): “Minimalist Market Design: A Framework for Economists with Policy Aspirations,” *Econometric Society Monographs Series*, forthcoming.
- YOKOI, Y. (2019): “Matroidal Choice Functions,” *SIAM Journal on Discrete Mathematics*, 33(3), 1712–1724.