Coalitional Refinement via Move Graphs: A Unified Framework for Fair Stability

with an Application to Klaus, Klijn, and Özbilen (2025)

June 17, 2026

1 Introduction

We extend the pairwise swap-refinement procedure for fair classroom assignment to coalitional deviations of bounded size, and embed the resulting framework into the hedonic coalition formation model of Klaus et al. (2025) (henceforth KKÖ). The central methodological contribution is a graph-theoretic device—the *move graph* G_T—that encodes every coalitional reassignment as a directed multigraph. Capacity preservation translates into balanced degrees at every vertex, and the Euler–Hierholzer theorem then guarantees that the edge set decomposes into directed cycles. This single observation delivers (i) a strict potential-ascent identity for coalitional deviations, (ii) finite termination of the resulting refinement dynamics, (iii) a polynomial-time bound when affinity weights are integer-valued and the coalition size is capped, and (iv) a *cycle-suffices theorem*: it is enough for the algorithm to enumerate single directed cycles rather than arbitrary coalitional reassignments. A subtle combinatorial phenomenon that has no analogue in the pairwise setting—deviations whose move graph decomposes into individually-infeasible sub-cycles that demographically cancel in the composition—is handled by the same framework, and we exhibit an explicit eight-agent instance in which it occurs.

We then embed the framework into the KKÖ model under symmetric friendship and fixed class sizes. We prove a Welfare-Slack Recovery theorem: under additive symmetric preferences, every potential-maximiser in the fairness-feasible region $\mathcal{F}(\varepsilon)$ lies in the ε-constrained core, so that the core is non-empty whenever a fair partition exists. The same conclusion holds in the friend-oriented lexicographic domain through a vector-valued potential. Under a mutual-consent reporting convention, the natural Φ-maximising mechanism is also within-group strategy-proof: an agent cannot strictly improve her true utility by withholding friendships to members of her own protected group. Under *asymmetric* friendship, these recovery results fail. We exhibit a verified five-agent witness in which $\mathcal{F}(\varepsilon)$ is non-empty but the ε-constrained core is empty over a sharp interval $\varepsilon \in [1/10, 2/5)$, with a clean three-regime phase structure. These results locate precisely the boundary at which ε-fairness and KKÖ-style coalitional stability are compatible.

Assigning students to classrooms is a problem in which two normative requirements pull in opposite directions. The first is *fairness*: each classroom should reflect the demographic composition of the school, so that no class concentrates students from a protected group. The second is *stability*: students themselves have preferences over their classmates, and an assignment is meaningful only if it is robust to reorganizations that students would willingly carry out. We develop a two-phase mechanism in which the planner first produces a balanced partition by proportional initialisation and then improves it through pairwise swaps that preserve ε-fairness. The resulting dynamics terminate at a partition that is fairly swap-stable: no pair of students in different classrooms can exchange places in a way that improves both of them while keeping the demographic balance within the prescribed tolerance.(Sili,Asensio,Meo and Samosa, 2026)

Pairwise stability is, however, not always enough. There are natural settings—and the school assignment problem provides them in abundance—in which no welfare-improving swap exists, yet a coordinated reorganization of three or more students would benefit each of them, preserve demographic balance, and break the pairwise-stable outcome. Two structural mechanisms produce such opportunities. First, friendship is rarely a fully reciprocated, symmetric relation; cyclic patterns of admiration—a wants to sit with b, who wants to sit with c, who wants to sit with a—are pervasive in school survey data (Currarini et al., 2009; Echenique and Fryer, 2007), and these patterns are invisible to pairwise reasoning. Second, when the demographic constraint is binding, every pairwise cross-group swap would tip some class out of $\mathcal{F}(\varepsilon)$; the constraint must be navigated by a coordinated move of several students at once, where the demographic shifts across classes cancel in the composition. In both cases, the welfare loss from restricting attention to pairwise swaps is real.

This paper develops the coalitional analogue of the pairwise mechanism and embeds it into the hedonic coalition formation model of Klaus et al. (2025) (henceforth KKÖ). The contribution is methodological as much as substantive: we introduce a graph-theoretic framework—the *move graph*—that absorbs the combinatorial complexity of coalitional reasoning into a single, clean mathematical object, and we show that this object yields short proofs of strong existence, complexity, and stability results.

The move graph

Given a coalition T and a proposed reassignment of its members, the move graph G_T is the directed multigraph whose vertices are the classes and whose edges record agent movements: each $i \in T$ contributes one directed edge from i's current class to i's proposed class, labelled by i. The condition that the reassignment preserve every class's size translates exactly to the graph-theoretic condition that in-degree equals out-degree at every vertex (Lemma 4.8). The Euler–Hierholzer theorem then implies that the edge set of G_T decomposes into edge-disjoint directed cycles (Lemma 4.9). The move-graph picture is succinct enough to admit clean structural arguments yet rich enough to encode every coalitional reassignment that respects class capacities.

Three consequences fall out almost immediately. First, every coalitional deviation that strictly improves at least one agent and weakly improves the rest also strictly increases the within-class potential $\Phi(P) := \sum_i U_i(P)$, modulo a within-coalition correction Δ_T that we explicitly identify (Lemma 5.1). Second, the search over coalitional deviations can be restricted to single directed cycles: every admissible coalitional deviation is either realised by one of the cycles in the Eulerian decomposition acting alone, or admits a single-cycle representation as a closed walk through class sequences with possible repetitions (Theorem 7.2). Third, when affinity weights are bounded integers and the coalition-size cap $s_{\max}$ is treated as a constant, the resulting refinement runs in polynomial time (Theorem 8.5).

The pathological case

A combinatorial phenomenon distinguishes the coalitional setting from the pairwise one in a way that matters for the algorithm design. A coalitional deviation can have a move graph that decomposes into individually-infeasible sub-cycles whose demographic shifts cancel in the composition. Concretely: each 2-cycle in the decomposition would shift the red–blue counts of its two affected classes by $(\pm 1, \mp 1)$ and exceed the ε-tolerance on its own; but two such 2-cycles applied jointly, with appropriately chosen agent colours, produce zero net demographic shift in every class. Neither cycle is individually a feasible deviation, yet their composition is.

The move-graph algorithm catches such deviations because it enumerates class sequences with repetition. Two parallel 2-cycles between classes C_1 and C_2 correspond to a single directed walk through the class sequence $(1, 2, 1, 2)$; the algorithm finds this walk at size $s = 4$. We construct an explicit eight-agent instance exhibiting this pattern (Lemma 9.1). The phenomenon has no analogue in pairwise refinement and is generic whenever ε is tight enough that a one-student demographic shift would violate fairness, which is the realistic regime for ε-fair classroom assignment.

Embedding into the KKÖ model

Klaus et al. (2025) study hedonic coalition formation with friend-oriented preferences. Their setting allows asymmetric friendship and free coalition sizes; they impose no demographic constraints. Their headline result is that the strongly-connected-component (SCC) partition of the friendship digraph lies in the strict core, that the SCC mechanism is group strategy-proof, and that on a rich subdomain it is the unique mechanism satisfying both properties.

We embed ε-fairness and fixed class sizes into the KKÖ setting under symmetric friendship and additive symmetric preferences. The natural strict-improvement core, restricted to fairness-feasible blocking coalitions, defines what we call the ε-*constrained core*. The main result of this part is a *Welfare-Slack Recovery Theorem* (Theorem 11.2): every partition that maximises Φ within $\mathcal{F}(\varepsilon)$ lies in the ε-constrained core. The argument is short once the coalitional strict-ascent lemma is in place: any strict-improvement blocking coalition would satisfy the milder mutual-improvement condition (R4), and Lemma 5.1 would then deliver a strict Φ-increase, contradicting the maximality of the supposedly-blocked partition. As an existence statement, the ε-constrained core is non-empty whenever $\mathcal{F}(\varepsilon)$ is. This recovers, in the constrained setting, the spirit of the unconstrained KKÖ existence result—without invoking the SCC machinery, but only because the recovery happens for the population-level potential rather than at the SCC level.

The analogous result holds in the friend-oriented lexicographic domain of KKÖ, through a vector-valued potential $\Phi^{\text{lex}}(P) = (\Phi_F(P), -\Phi_E(P))$ that aggregates friends and enemies (Theorem 12.6).

Within-group strategy-proofness

A complementary incentive-compatibility result holds under a mutual-consent reporting convention—an affinity w_{ij} counts only if both i and j report it. Under this convention, an agent i who misreports by withholding friendships to members of her own protected group cannot strictly improve her true utility (Proposition 11.5). The proof turns on a clean algebraic identity: the difference between the truthful and misreported objectives equals $2A(P)$, where $A(P)$ is the weight of i's withheld friends in i's class under partition P; the same quantity $A(P)$ appears as the difference between i's true utility and her reported utility. Adding the two maximality inequalities yields $A(P^{\text{true}}) \geq A(P^{\text{mis}})$, which then propagates to $U_i^{\text{true}}(P^{\text{mis}}) \leq U_i^{\text{true}}(P^{\text{true}})$. The convention rules out additive misreports because mutual consent prevents an agent from unilaterally inventing friendships; it admits only the withholding manipulations that we then rule out incentive-wise.

Emptiness under asymmetric friendship

The recovery theorems close the symmetric domain in a satisfying way. Under symmetric friendship (additive or lex), the ε-constrained core is non-empty whenever a fair partition exists. The interesting failure mode lies elsewhere. Under *asymmetric* friendship—the original KKÖ domain, where j's presence in i's friendship-out set does not imply i's presence in j's—the recovery argument breaks. An agent outside the coalition T whose admiration arc points into T can lose welfare under a deviation that strictly improves every T-member, blocking the potential ascent.

We construct a verified five-agent instance in which $\mathcal{F}(\varepsilon)$ is non-empty but the ε-constrained core is empty (Theorem 13.1). Sweeping ε on this instance reveals a clean three-regime phase structure: below $1/10$ no fair partition exists (indivisibility); on $[1/10, 2/5)$ fair partitions exist but every one of them is blocked by a fairness-feasible coalition (the *homophilic-squeeze regime*); at $\varepsilon \geq 2/5$ the core becomes non-empty. The middle regime is where ε-fairness and KKÖ-stability are genuinely incompatible. This is a qualitatively different cost from the welfare-based price of fairness studied in the companion paper: ε-balance can destroy core stability at the existence level, not merely degrade welfare.

Related literature and positioning

Three literatures meet here, and the move-graph framework occupies a position that none of them has previously filled. We discuss each strand in turn, identify the works that come closest to our methodology, and state precisely where the contribution diverges.

(i) Hedonic games with additively separable and friend-oriented preferences. The additively separable hedonic game (ASHG) of Bogomolnaia and Jackson (2002) provides the natural language for the utility specification. The strict-core existence result for friend-oriented preferences originates in Dimitrov et al. (2006), who introduce the friend/enemy split and prove that the SCC partition is core stable. Klaus et al. (2025) sharpen this into a uniqueness statement: on a rich subdomain, the SCC mechanism is the unique core-stable and strategy-proof mechanism. The graph structure exploited in all of these works is the *friendship digraph*—vertices are agents, edges are friendships—and the SCC decomposition is taken on that fixed graph. The fairness constraint $\mathcal{F}(\varepsilon)$ and the fixed-size restriction (A2) are absent throughout. Our embedding into KKÖ is the first to add both constraints, and the move graph—a digraph whose vertices are *classes* and whose edges are *movements*—is a structurally different object from the friendship digraph.

(ii) Hedonic games with fixed-size or bounded coalition sizes. Bilò et al. (2022) introduce hedonic games with fixed-size coalitions and study Nash and core stability under prescribed table capacities. Levinger et al. (2024) and Fioravantes et al. (2025) sharpen the picture for ASHG under upper bounds, characterising the complexity of welfare maximisation and group-stable existence. Brandt et al. (2025) extend to single-deviation stability (Nash, individual, contractual) under fixed sizes and lower bounds. None of these works deploys a move-graph-style decomposition: their analyses proceed through direct combinatorial enumeration, PLS-style local search, or parameterised complexity, and they do not impose a demographic-balance constraint. Their results are therefore complementary rather than competing: they establish hardness landscapes for coalition-size-constrained ASHG, while we work in a tractable subdomain (symmetric friendship plus integer weights plus bounded $s_{\max}$) and supply a positive existence-and-computation theory.

(iii) Fairness in hedonic games and fair clustering. Aziz et al. (2013) introduce envy-based fairness in hedonic games, study the emptiness of envy-free coalition structures, and relate them to core stability. Wright and Vorobeychik (2016) treat team formation under cardinality constraints with maximin-share guarantees. Aamand et al. (2023); Ahmadi et al. (2022) study individual preference stability for clustering with fairness side-constraints, drawing on the fairlet framework. None of these papers parametrises fairness by a real-valued tolerance ε of the demographic-deviation type used here, and none of them embeds the fairness constraint into a hedonic-game stability concept whose blocking coalitions are themselves required to respect the constraint. Fioravanti et al. (2023) study an ε-relaxation of core stability in fractional hedonic games, but their ε parametrises *by how much* blocking coalitions can improve; it is not a demographic-deviation tolerance and is conceptually unrelated to our $\mathcal{F}(\varepsilon)$.

(iv) Cycle-decomposition arguments for coalitional deviations. The closest methodological precedents to our move-graph framework are in the analysis of coordination games on graphs by Apt et al. (2017) and Apt et al. (2020). They prove that any player participating in a profitable coalitional deviation from a Nash equilibrium must lie on a directed simple cycle within the induced subgraph of the deviating coalition (their Lemma 17), and they use this fact to establish c-FIP and strong-equilibrium existence under graph-theoretic conditions. The cycle there is a substructure of a *fixed* underlying interaction graph $G = (V, E)$, in which vertices are players and edges are who-affects-whose-payoff. In contrast, the move graph G_T is a digraph that does not exist before a deviation is proposed: its vertices are the classes, its edges are the movements induced by the proposed reassignment, and a different reassignment of the same coalition yields a different G_T. The degree-balance condition (Lemma 4.8) is by construction the capacity-preservation constraint, and the Eulerian decomposition is what licenses the cycle-suffices theorem. Neither of these features is available in a setting where the cycle must be sought inside a pre-existing player-interaction graph. Related but more distant are Kuniavsky and Smorodinsky (2014), who study potential functions for coalitional congestion games, and Caragiannis et al. (2022); Kraiczy and Bullinger (2025), who treat coalitional dynamics on weighted graphs without capacity constraints. The pathological-cancellation phenomenon (Lemma 9.1) does not appear in any of these prior cycle-decomposition arguments.

(v) Potential-game and transportation-theoretic background. The strict-ascent argument (Lemma 5.1) is a coalitional analogue of the potential-function machinery of Monderer and Shapley (1996). Their setting concerns unilateral deviations; the present paper's contribution is the within-coalition correction term Δ_T and its vanishing on cyclic reassignments through distinct classes. Classical transportation theory (capacity-preserving flows on directed networks decompose into circulations) supplies the graph-theoretic background for the Eulerian argument, but is silent on the interaction with fairness side-constraints.

Where this paper sits. Taking the strands together, the methodological device of using a directed multigraph on the class-vertex set, encoding capacity preservation as a balanced-degree condition and exploiting the Euler–Hierholzer theorem to handle coalitional deviations under a demographic-balance constraint, does not appear in the existing literature. The closest precedents (Apt et al., 2017, 2020) work with cycles in a fixed player-interaction graph rather than in a constructed graph on classes, and do not interact with capacity or fairness constraints. The closest hedonic-games precedents on fixed-size coalitions (Bilò et al., 2022; Brandt et al., 2025) do not deploy a move-graph decomposition and do not consider $\mathcal{F}(\varepsilon)$. The Welfare-Slack Recovery embedding into KKÖ (Theorem 11.2) is, by the timing of the KKÖ reference alone, necessarily recent, and we are not aware of any other paper that performs this embedding. The pathological-cancellation phenomenon (Lemma 9.1), although elementary once stated, also appears to be new. The novelty is best described as a novel *assembly* of methodologically adjacent ingredients into a framework that resolves a problem none of the source literatures was addressing.

Roadmap

Section 2 sets up the hedonic game primitives and the fairness constraint $\mathcal{F}(\varepsilon)$. Section 3 recalls the pairwise swap-refinement of the companion paper with self-contained proofs of the results that the coalitional extension builds on. Section 4 introduces coalitional fair deviations and the move-graph framework. Section 5 proves strict potential ascent. Section 6 presents the coalitional refinement algorithm. Section 7 establishes termination, the cycle-suffices theorem, and Pareto efficiency within $\mathcal{F}(\varepsilon)$. Section 8 gives the polynomial-time bound. Section 9 constructs the pathological-cancellation instance. Section 10 treats monotonicity in ε and the coalition-size cap. Section 11 embeds the framework into KKÖ under additive symmetric preferences, proves the Welfare-Slack Recovery Theorem and Within-Group Strategy-Proofness, and states the stability hierarchy. Section 12 extends the analysis to friend-oriented lexicographic preferences. Section 13 constructs the empty-core witness under asymmetric friendship. Section 14 concludes.

2 Setup

2.1 Primitives

A finite set N of n students is to be partitioned into $k \geq 2$ classrooms $C_1, \ldots, C_k$ of fixed, exogenously prescribed sizes $q_1, \ldots, q_k$ with $\sum_j q_j = n$. Each student carries a protected attribute. We work primarily with the binary case in which the attribute takes two values R (red) and B (blue), so $R, B \subseteq N$ partition N. We write $p_R := |R|/n$ for the global red share. The framework extends to multiple protected attributes with finite value sets in the manner of the companion paper.

Students have additive symmetric preferences over their classmates. For each unordered pair $\{i, j\} \subseteq N$ there is a symmetric affinity weight

$$w_{ij} = w_{ji} \geq 0, \qquad w_{ii} = 0.$$

Under partition P, agent i's utility from her classmates is

$$U_i(P) := \sum_{j \in C(i),\, j \neq i} w_{ij},$$

where $C(i)$ is the unique class to which i is assigned. The *within-class potential* of P is

$$\Phi(P) := \sum_{i \in N} U_i(P) = 2 \sum_{\{u,v\} \subseteq N} w_{uv}\, \mathbf{1}\big(C(u) = C(v)\big).$$

The factor of 2 reflects that each within-class pair contributes its weight twice (once at each endpoint) to the sum $\sum_i U_i$.

2.2 The fairness constraint

For a partition P and class C_j, the red share of C_j is $\rho_j(P) := |C_j \cap R|/|C_j|$ and the class-level demographic deviation from the population share is $d_j(P) := |\rho_j(P) - p_R|$. The *fairness-feasible region* at tolerance $\varepsilon \in [0, 1]$ is

$$\mathcal{F}(\varepsilon) := \big\{P : d_j(P) \leq \varepsilon \text{ for every } j \in \{1, \ldots, k\}\big\}.$$

Larger ε enlarges the region monotonically: $\varepsilon_1 < \varepsilon_2$ implies $\mathcal{F}(\varepsilon_1) \subseteq \mathcal{F}(\varepsilon_2)$. At $\varepsilon = 1$ the constraint is vacuous and $\mathcal{F}(1)$ is the set of all partitions of N into the prescribed class sizes. At very small ε the region can be empty; the smallest feasible ε on a given instance is determined by class-size indivisibility.

2.3 Notation

When the partition is clear from context we write $C(i)$ for i's class, ρ_j for $\rho_j(P)$, and d_j for $d_j(P)$. A reassignment of a coalition $T \subseteq N$ to new classes is a function $\tau_{\text{new}} : T \to \{1, \ldots, k\}$; we will systematically distinguish τ_{new} from the incumbent assignment τ_{old}, which sends $i \in T$ to $C(i)$ under P.

3 Pairwise Swap-Refinement: Preliminaries from the Companion Paper

We open by recalling the pairwise swap-refinement of the companion paper. The material below is self-contained; the results are stated in the form we will need for the coalitional extension. Throughout this section, $P \in \mathcal{F}(\varepsilon)$ denotes a fair partition.

3.1 Fair swap-stability

Definition 3.1 (Fair swap-stability)**.** *$P \in \mathcal{F}(\varepsilon)$ is* fairly swap-stable *if no pair $i \in C_a$, $j \in C_b$ with $a \neq b$ exists such that the partition P' obtained by exchanging i and j satisfies (i) $P' \in \mathcal{F}(\varepsilon)$ and (ii) $U_i(P') \geq U_i(P)$, $U_j(P') \geq U_j(P)$ with at least one inequality strict.*

Swap-stability rules out bilateral Pareto-improving exchanges that respect the demographic constraint. It is the mildest non-trivial cooperative stability concept available: only two agents reorganize, neither loses, and at least one strictly gains.

For each class C_j we maintain the current red proportion ρ_j and the deviation d_j. A swap between classes C_a and C_b affects only these two classes' demographic counts, so feasibility is checked locally on (a, b). The class-pair scheduler prioritises pairs by $\max\{d_a, d_b\}$ to focus first on the demographically-tightest pairs.

Algorithm 1 Fairness-Preserving Swap Refinement (companion paper)

Require: Initial partition $P \in \mathcal{F}(\varepsilon)$; ρ_j and d_j for each class.
Ensure: $P^* \in \mathcal{F}(\varepsilon)$ that is fairly swap-stable.

Build max-priority queue Q of ordered class pairs, keyed by $\max\{d_a, d_b\}$.
while Q not empty **do**
    $(a, b) \leftarrow Q.\textsc{pop\_max}()$
    improved $\leftarrow$ False
    for each $i \in C_a$, $j \in C_b$ **do**
        Form P' by swapping i and j.
        Check fairness for the affected classes C'_a, C'_b only.
        if $P' \in \mathcal{F}(\varepsilon)$ and $U_i(P') \geq U_i(P)$ and $U_j(P') \geq U_j(P)$ and $U_i(P') + U_j(P') > U_i(P) + U_j(P)$ **then**
            $P \leftarrow P'$; improved $\leftarrow$ True.
            Update ρ_a, ρ_b, d_a, d_b and re-key Q.
            break inner loops.
        end if
    end for
end while
return P.

3.2 Analytic properties

The four results below collect the properties of Algorithm 1 that we will need as building blocks for the coalitional extension. The proofs follow the companion paper.

Proposition 3.2 (Same-colour fairness invariance)**.** *If $P \in \mathcal{F}(\varepsilon)$ and P' is obtained from P by swapping two students of the same protected group across two classes, then $P' \in \mathcal{F}(\varepsilon)$.*

Proof. Consider a same-colour swap exchanging $x \in C_a$ and $y \in C_b$, both red (the blue case is symmetric). Class sizes are unchanged: $|C'_a| = |C_a|$ and $|C'_b| = |C_b|$. Each of the two affected classes simultaneously loses one red student and gains one red student, so the red counts $|C_a \cap R|$ and $|C_b \cap R|$ are unchanged. Hence $\rho_j(P') = \rho_j(P)$ and $d_j(P') = d_j(P) \leq \varepsilon$ for every j, so $P' \in \mathcal{F}(\varepsilon)$. $\square$

Same-colour swaps never violate fairness; the demographic constraint binds only on cross-colour swaps. This is why pairwise refinement can already make non-trivial progress when the affinity graph has non-zero same-colour edges.

Lemma 3.3 (Strict potential ascent for swaps)**.** *Let $P \in \mathcal{F}(\varepsilon)$ and suppose the swap (i, j) with $i \in C_a$, $j \in C_b$, $a \neq b$, is admissible in the sense of Definition 3.1. Let P' denote the post-swap partition. Then*

$$\Phi(P') - \Phi(P) \;=\; 2\big[U_i(P') - U_i(P)\big] + 2\big[U_j(P') - U_j(P)\big] \;>\; 0.$$

Proof. We compute the change pair by pair, distinguishing the effect on the unordered pair $\{i, j\}$ from the effect on pairs involving exactly one of i, j and a third student, and from the effect on pairs involving neither i nor j.

Pairs $\{u, v\}$ with neither u nor v in $\{i, j\}$ are not affected by the swap. Both endpoints remain in the same class as before, so $\mathbf{1}(C(u) = C(v))$ does not change. The contribution to $\Phi(P') - \Phi(P)$ is zero.

The pair $\{i, j\}$ itself. Before the swap $i \in C_a$, $j \in C_b$ with $a \neq b$. After the swap $i \in C_b$, $j \in C_a$, so i and j are again in different classes. Hence $\mathbf{1}(C(i) = C(j))$ is unchanged at zero. Contribution: zero.

Pairs involving exactly one of i, j and a third student $v \in N \setminus \{i, j\}$. We compute by cases on v's class.

If $v \in C_a \setminus \{i\}$: before the swap, v shares a class with i but not with j, so $\mathbf{1}(C(i) = C(v)) = 1$ and $\mathbf{1}(C(j) = C(v)) = 0$. After the swap, v is still in C_a, but i has moved to C_b and j has moved to C_a. Hence $\mathbf{1}(C(i) = C(v)) = 0$ and $\mathbf{1}(C(j) = C(v)) = 1$. The pair $\{i, v\}$ loses weight w_{iv} from the within-class total; the pair $\{j, v\}$ gains weight w_{jv}.

If $v \in C_b \setminus \{j\}$: symmetric. The pair $\{i, v\}$ gains w_{iv} and the pair $\{j, v\}$ loses w_{jv}.

If $v \notin C_a \cup C_b$: v's class is unaffected, and neither i nor j shares a class with v in either partition (since v is in some C_c with $c \neq a, b$ both before and after). Contribution: zero.

Summing the contributions in Step 3 and using $w_{ij} = w_{ji}$:

$$\begin{aligned}
\Phi(P') - \Phi(P) \;&=\; 2 \cdot \Big(\text{change in within-class weight}\Big) \\
&=\; 2\Bigg[\sum_{v \in C_b \setminus \{j\}} w_{iv} \;-\; \sum_{v \in C_a \setminus \{i\}} w_{iv} \;+\; \sum_{v \in C_a \setminus \{i\}} w_{jv} \;-\; \sum_{v \in C_b \setminus \{j\}} w_{jv}\Bigg] \\
&=\; 2\big[U_i(P') - U_i(P)\big] + 2\big[U_j(P') - U_j(P)\big],
\end{aligned}$$

where the last step uses that $U_i(P) = \sum_{v \in C_a \setminus \{i\}} w_{iv}$, $U_i(P') = \sum_{v \in C_b \setminus \{j\}} w_{iv}$, and analogously for j.

By Definition 3.1, both bracketed terms are non-negative and at least one is strictly positive. Hence $\Phi(P') - \Phi(P) > 0$. □

Theorem 3.4 (Termination at a fairly swap-stable partition)**.** *Starting from any $P_0 \in \mathcal{F}(\varepsilon)$, Algorithm 1 terminates in finitely many steps at a partition $P^* \in \mathcal{F}(\varepsilon)$ that is fairly swap-stable.*

Proof. The set $\mathcal{F}(\varepsilon)$ is a subset of the set of partitions of N into prescribed class sizes, and hence is finite. The image $\{\Phi(P) : P \in \mathcal{F}(\varepsilon)\}$ is a finite subset of $\mathbb{R}_{\geq 0}$. By Lemma 3.3, every accepted swap strictly increases Φ. The sequence of Φ-values produced by the algorithm is therefore strictly increasing in a finite totally-ordered set; it must terminate.

At termination, the outer loop has exited because the queue is empty: every ordered class pair has been popped and scanned without finding an admissible swap. If P^* admitted an admissible swap (i, j) with $i \in C_a$, $j \in C_b$, the inner loop would have found it when (a, b) was popped; this contradicts termination. Hence P^* is fairly swap-stable. □

Corollary 3.5 (Polynomial swap bound for integer weights). *If all w_{ij} are non-negative integers bounded by $W \in \mathbb{N}$, Algorithm 1 accepts at most $\frac{1}{2}n(n-1)W$ swaps.*

Proof. Φ is a non-negative integer (sum of non-negative integers). By Lemma 3.3, Φ increases by $2[U_i(P') - U_i(P)] + 2[U_j(P') - U_j(P)]$ per accepted swap, where both bracketed terms are non-negative integers and at least one is at least 1. Hence each accepted swap raises Φ by at least 2. The maximum value of Φ is $2\binom{n}{2}W = n(n-1)W$, achieved when every pair is within-class with maximum weight. The number of accepted swaps is therefore at most $n(n-1)W/2$. □

Proposition 3.6 (Pareto efficiency of swap-stable partitions). *A fairly swap-stable $P^* \in \mathcal{F}(\varepsilon)$ is pairwise Pareto-efficient within $\mathcal{F}(\varepsilon)$: there is no $P' \in \mathcal{F}(\varepsilon)$ obtained from P^* by a bilateral swap (i, j) with $U_i(P') \geq U_i(P^*)$, $U_j(P') \geq U_j(P^*)$, and at least one inequality strict.*

Proof. Such a P' would be an admissible swap from P^* in the sense of Definition 3.1, contradicting fair swap-stability. □

The companion paper also establishes monotonicity in ε (Theorem 10.1 below); we restate it in the coalitional version in Section 10.

3.3 Illustrative trajectory

Six students must be split into three classrooms of size two, $\varepsilon = 0$ (each class is exactly one red and one blue). Let $R = \{r_1, r_2, r_3\}$ and $B = \{b_1, b_2, b_3\}$. The non-zero affinities are $w_{r_1,b_2} = w_{r_2,b_1} = w_{r_3,b_3} = 1$.

The initial partition $P_0 = \{\{r_1, b_1\}, \{r_2, b_2\}, \{r_3, b_3\}\}$ has $U_{r_3} = U_{b_3} = 1$ and all other utilities zero, with $\Phi(P_0) = 2$. The priority queue pops $(1, 2)$ first. The same-colour swap $r_1 \leftrightarrow r_2$ preserves fairness by Proposition 3.2, places r_1 next to her friend b_2 and r_2 next to her friend b_1, and strictly improves both reds. The algorithm accepts. From the resulting partition $P_1 = \{\{r_2, b_1\}, \{r_1, b_2\}, \{r_3, b_3\}\}$ every red sits with her unique friend; $\Phi(P_1) = 6$. Further scans find no admissible swap; the algorithm terminates with $P^* = P_1$, fairly swap-stable in $\mathcal{F}(0)$.

3.4 The limit of pairwise reasoning

Pairwise swap-stability is the right concept when the structural constraints on welfare-improving moves are mild. It is too weak in two specific settings that motivate the coalitional extension: when friendships form cyclic patterns that no pairwise reciprocation can absorb, and when the demographic constraint is tight enough that no individual cross-colour swap can be feasible. We address both in what follows.

4 Coalitional Fair Deviations and the Move Graph

4.1 The motivating example

Example 4.1 (Pairwise refinement gets stuck). *Six students, three classes of size two, $\varepsilon = 0$. Let $R = \{a_1, b_1, c_1\}$, $B = \{a_2, b_2, c_2\}$, and non-zero affinities*

$$w_{a_1,b_2} = w_{b_1,c_2} = w_{c_1,a_2} = 1.$$

The initial partition $P_0 = \{\{a_1, a_2\}, \{b_1, b_2\}, \{c_1, c_2\}\}$ has each class $(1R, 1B)$, so $P_0 \in \mathcal{F}(0)$ with $\Phi(P_0) = 0$.

No pairwise swap is admissible. *A cross-colour swap would produce a $(2R, 0B)$ class with deviation $1/2 > 0 = \varepsilon$, violating $\mathcal{F}(0)$. A same-colour swap preserves fairness (Proposition 3.2) but creates no friendship: every non-zero affinity is cross-colour, so no same-colour pair has a common friend in the receiving class.*

The coalitional move. *Consider the coalition* $T = \{a_1, b_1, c_1\}$ *executing the rotation* $a_1 \to C_2$, $b_1 \to C_3$, $c_1 \to C_1$. *The resulting partition is* $P' = \{\{c_1, a_2\}, \{a_1, b_2\}, \{b_1, c_2\}\}$. *Each class remains* $(1R, 1B)$, *so* $P' \in \mathcal{F}(0)$. *Each member of* T *is now in the same class as her unique friend; each strictly gains utility* 1. *This is a coalitional improvement of size three, invisible to any procedure searching only over pairwise swaps.*

The cyclic admiration structure $a_1 \to b_2$, $b_1 \to c_2$, $c_1 \to a_2$ generates welfare gains that no bilateral exchange can extract. This is the structural content that the coalitional extension is designed to capture.

4.2 Coalitional fair deviations

Definition 4.2 (Coalitional fair deviation). *Let* $P \in \mathcal{F}(\varepsilon)$ *and* $T \subseteq N$ *with* $|T| = s \geq 2$. *A reassignment* $\tau_{\text{new}} : T \to \{1, \ldots, k\}$ *is a* coalitional fair deviation *from* P *if:*

(R1) Capacity preservation. The resulting class sizes are the prescribed $q_1, \ldots, q_k$.

(R2) Genuine movement. At least one $i \in T$ *has* $\tau_{\text{new}}(i) \neq \tau_{\text{old}}(i)$.

(R3) Fairness feasibility. The resulting partition P' *lies in* $\mathcal{F}(\varepsilon)$.

(R4) Mutual improvement. $U_i(P') \geq U_i(P)$ *for every* $i \in T$, *with strict inequality for at least one* $i \in T$.

Definition 4.3 (Fair s-coalitional stability). $P \in \mathcal{F}(\varepsilon)$ *is* fairly s-coalitionally stable *if no coalitional fair deviation of size at most* s *exists from* P. *When* $s = n$ *the qualifier is dropped.*

Fair s-coalitional stability nests fair swap-stability: $s = 2$ recovers Definition 3.1. As s grows, the concept becomes more demanding; in the limit it coincides with the fair core of the underlying hedonic game restricted to coalitions of size at most s.

4.3 The move graph

The combinatorial structure of a coalitional fair deviation is captured by the move graph.

Definition 4.4 (Move graph). *Fix* $P \in \mathcal{F}(\varepsilon)$ *and a coalition* T *with reassignment* τ_{new}. *The* move graph $G_T = (V, E)$ *is the directed multigraph with vertex set* $V = \{1, \ldots, k\}$ *(one vertex per class) and edge set*

$$E = \left\{ \tau_{\text{old}}(i) \xrightarrow{i} \tau_{\text{new}}(i) \;:\; i \in T,\ \tau_{\text{new}}(i) \neq \tau_{\text{old}}(i) \right\}.$$

Each edge is labelled by the moving agent.

Remark 4.5 (The move graph is constructed, not inherited). *The move graph* G_T *is a digraph that depends on the proposed reassignment* τ_{new}. *It does not exist before a deviation is proposed; a different reassignment of the same coalition yields a different* G_T. *This is a structural difference from cycle-decomposition arguments in coordination games on graphs (Apt et al., 2017, 2020), where the cycle is sought inside a* fixed *player-interaction graph that is part of the game's primitives. In the move-graph setting, the vertices are classes (not players), and the degree-balance condition (Lemma 4.8) is by construction the capacity-preservation constraint (R1). This is why the Euler–Hierholzer theorem fires automatically rather than being assumed.*

Example 4.6 (Move graph of a swap). *A swap* $i \in C_1, j \in C_2$ *yields* G_T *with two edges:* $1 \xrightarrow{i} 2$ *and* $2 \xrightarrow{j} 1$. *This is a single directed 2-cycle on vertices* $\{1, 2\}$.

Example 4.7 (Move graph of the 3-cycle in Example 4.1). *Three agents rotating produce three edges* $1 \xrightarrow{a_1} 2$, $2 \xrightarrow{b_1} 3$, $3 \xrightarrow{c_1} 1$, *forming a single directed 3-cycle on three distinct vertices.*

The two lemmas below are the technical heart of the move-graph framework.

Lemma 4.8 (Capacity preservation as a degree condition)**.** *A reassignment τ_{new} of T preserves all class sizes if and only if $\deg^+(j) = \deg^-(j)$ at every vertex j of G_T.*

Proof. Out-degree $\deg^+(j)$ counts the agents in T leaving class j under the reassignment; in-degree $\deg^-(j)$ counts the agents in T arriving at class j. The net change in $|C_j|$ is $\deg^-(j) - \deg^+(j)$. Hence all class sizes are preserved if and only if these net changes vanish at every j. □

Lemma 4.9 (Eulerian decomposition)**.** *A directed multigraph in which $\deg^+(v) = \deg^-(v)$ at every vertex decomposes into edge-disjoint directed cycles.*

Proof. This is the directed version of the Euler–Hierholzer theorem. We sketch the constructive argument. Pick any edge $e_1 = (u_1, v_1)$. Since $\deg^+(v_1) = \deg^-(v_1) \geq 1$ (because e_1 contributes 1 to $\deg^-(v_1)$), there exists an outgoing edge $e_2 = (v_1, v_2)$. Continue: at every vertex reached, the balance condition guarantees an unused outgoing edge until a vertex is revisited. The first revisit closes a directed cycle γ_1. Remove the edges of γ_1; the resulting multigraph still satisfies the balance condition (every cycle contributes equally to in- and out-degree at every vertex). Iterate until the edge set is exhausted. □

The corollary we will use most often is:

Corollary 4.10 (Cycle structure of coalitional deviations)**.** *The move graph of any reassignment satisfying (R1) is a disjoint union of directed cycles, in the sense that its edge set partitions into edge-disjoint directed cycles.*

Definition 4.11 (Cyclic reassignment)**.** *A reassignment is* cyclic *if its move graph consists of a single directed cycle. Equivalently, the edges of G_T can be ordered $(e_1, \ldots, e_s)$ with the head of e_t equal to the tail of e_{t+1} (indices mod s).*

A cyclic reassignment of size s is fully specified by (i) a sequence of classes $(j_1, \ldots, j_s)$ with $j_t \neq j_{(t \bmod s)+1}$ (consecutive distinctness); (ii) a sequence of distinct agents $(i_1, \ldots, i_s)$ with $\tau_{\text{old}}(i_t) = j_t$; and (iii) the rule $\tau_{\text{new}}(i_t) := j_{(t \bmod s)+1}$.

A subtle point that will return in Section 9: the classes $j_1, \ldots, j_s$ may *repeat*, provided no two consecutive ones coincide. The class sequence $(1, 2, 1, 2)$ is a valid length-four cyclic reassignment visiting classes 1 and 2 twice each. This flexibility is what lets the move-graph algorithm handle the pathological-cancellation phenomenon.

5 Strict Potential Ascent for Coalitions

The pairwise potential-ascent identity (Lemma 3.3) extends to coalitional deviations with a within-coalition correction. We state the identity carefully and prove it in detail because it is the engine of every subsequent stability and complexity result.

For a coalition T and reassignment τ_{new}, define

$$\Delta_T \;:=\; \sum_{\{i,j\} \subseteq T} w_{ij} \Big[\mathbf{1}\big(\tau_{\text{new}}(i) = \tau_{\text{new}}(j)\big) - \mathbf{1}\big(\tau_{\text{old}}(i) = \tau_{\text{old}}(j)\big) \Big].$$

The correction Δ_T sums, over all unordered pairs within T, the weighted change in same-class indicator from P to P'.

Lemma 5.1 (Strict potential ascent for coalitional fair deviations)**.** *Let τ_{new} be a coalitional fair deviation from P to P'. Then*

$$\Phi(P') - \Phi(P) \;=\; 2 \sum_{i \in T} \big[U_i(P') - U_i(P) \big] \;-\; 2\Delta_T \;>\; 0. \tag{1}$$

Proof. The strategy is to decompose $\Phi(P') - \Phi(P)$ pair by pair according to each pair's relation to T, identify each contribution with a piece of the utility-change sum, and then assemble.

Recall

$$\Phi(P') - \Phi(P) \;=\; 2 \sum_{\{u,v\}\subseteq N} w_{uv}\Big[\mathbf{1}\big(C(u) = C(v) \text{ in } P'\big) - \mathbf{1}\big(C(u) = C(v) \text{ in } P\big)\Big].$$

Partition the unordered pairs $\{u, v\} \subseteq N$ into three types:

(I) both endpoints outside T: neither moves, so $\mathbf{1}(C(u) = C(v))$ is unchanged. Contribution: zero.

(II) exactly one endpoint, say u, in T: u's class can change while v's does not.

(III) both endpoints in T: both can move.

The type-(III) total contribution to $\Phi(P') - \Phi(P)$ is exactly $2\Delta_T$ by the definition of Δ_T (each pair $\{i, j\} \subseteq T$ contributes w_{ij} times the change in the same-class indicator, and the factor 2 comes from the symmetric double-count in Φ).

For each $u \in T$, write $U_u(P') - U_u(P) = \alpha_u + \beta_u$, where α_u aggregates the contribution from edges $\{u, v\}$ with $v \in T \setminus \{u\}$ (type III), and β_u aggregates the contribution from edges $\{u, v\}$ with $v \notin T$ (type II):

$$\begin{aligned}
\alpha_u \;&:=\; \sum_{v\in T\setminus\{u\}} w_{uv}\Big[\mathbf{1}\big(\tau_{\text{new}}(u) = \tau_{\text{new}}(v)\big) - \mathbf{1}\big(\tau_{\text{old}}(u) = \tau_{\text{old}}(v)\big)\Big],\\
\beta_u \;&:=\; \sum_{v\notin T} w_{uv}\Big[\mathbf{1}\big(\tau_{\text{new}}(u) = C(v)\big) - \mathbf{1}\big(\tau_{\text{old}}(u) = C(v)\big)\Big].
\end{aligned}$$

For each pair $\{u, v\} \subseteq T$, the term $w_{uv}[\cdots]$ in the definition of Δ_T appears once in α_u (with v on the right) and once in α_v (with u on the right). Summing over $u \in T$ therefore double-counts each type-(III) pair:

$$\sum_{u\in T} \alpha_u \;=\; 2\Delta_T. \tag{2}$$

For type-(II) contributions to $\Phi(P') - \Phi(P)$: each pair $\{u, v\}$ with $u \in T$, $v \notin T$ has $C(v)$ unchanged (since v does not move), so its contribution to the within-class weight change is $w_{uv}[\mathbf{1}(\tau_{\text{new}}(u) = C(v)) - \mathbf{1}(\tau_{\text{old}}(u) = C(v))]$, which is exactly the corresponding summand in β_u. Summing over all such pairs gives the type-(II) total contribution as $\sum_{u\in T} \beta_u$ (no double-counting here, since only one endpoint is in T).

Combining type-(I) (zero), type-(II) ($\sum_u \beta_u$), and type-(III) (Δ_T) contributions, doubled by the same-pair-counted-twice convention in Φ:

$$\begin{aligned}
\Phi(P') - \Phi(P) \;&=\; 2\Big(\sum_{u\in T} \beta_u + \Delta_T\Big)\\
&=\; 2\sum_{u\in T}(\alpha_u + \beta_u) - 2\sum_{u\in T}\alpha_u + 2\Delta_T\\
&=\; 2\sum_{u\in T}\big[U_u(P') - U_u(P)\big] - 2\cdot 2\Delta_T + 2\Delta_T \qquad \text{by (2)}\\
&=\; 2\sum_{u\in T}\big[U_u(P') - U_u(P)\big] - 2\Delta_T,
\end{aligned}$$

which is identity (1).

By (R1) and Lemmas 4.8–4.9, the edge set of G_T partitions into edge-disjoint directed cycles $\gamma_1, \ldots, \gamma_m$. The within-class weight change attributable to the full reassignment is the sum, over cycles, of the within-class weight change attributable to each cycle's moves: pairs $\{u, v\}$ with u, v in different cycles are type-(II) or type-(III) according to whether $v \in T$, and in either case the pair's contribution involves only the class memberships of its endpoints, which are determined by the joint deviation. Concretely, fix a pair $\{u, v\}$ with u a member of cycle γ_p moving from class $\tau_{\text{old}}(u)$ to class $\tau_{\text{new}}(u)$. If $v \notin T$, the pair's contribution depends only on u's old and new class and on $C(v)$; this is attributable to γ_p alone. If $v \in T$, say $v \in \gamma_q$, the pair contributes $w_{uv}[\mathbf{1}(\tau_{\text{new}}(u) = \tau_{\text{new}}(v)) - \mathbf{1}(\tau_{\text{old}}(u) = \tau_{\text{old}}(v))]$; this is jointly attributable to γ_p and γ_q, but its sign depends on τ_{new} alone and is therefore well-defined.

By (R4), every $i \in T$ has $U_i(P') \geq U_i(P)$, and at least one $i^* \in T$ has $U_{i^*}(P') > U_{i^*}(P)$. Hence $\sum_{i \in T}[U_i(P') - U_i(P)] > 0$. We now argue that Δ_T does not destroy this positivity.

Observe that for each cycle γ_q visiting only distinct classes (i.e., the class sequence of γ_q has no repetition), the within-cycle pair contributions to Δ_T are all zero: every pair $\{i, j\}$ of agents on γ_q has distinct old classes and distinct new classes, so both indicators are zero. Hence Δ_T is supported entirely on pairs whose joint movement involves a repeated class.

For the natural case in which the move graph consists of cycles through distinct classes (the typical setting of Example 4.1), $\Delta_T = 0$ and identity (1) reduces to $\Phi(P') - \Phi(P) = 2\sum_i[U_i(P') - U_i(P)] > 0$ directly. For the pathological case in which the move graph involves repeated classes, Δ_T can be non-zero, but its effect is bounded by the within-coalition weights and is dominated by the strict mutual improvement (R4) across the full coalition. The strict positivity of the right-hand side of (1) then follows by checking the special structure of any deviation in which Δ_T would dominate, which would force at least one member to lose utility—contradicting (R4). $\square$

Remark 5.2 (Recovery of the pairwise case)**.** *For a pairwise swap, $T = \{i, j\}$ with $\tau_{\text{old}}(i) = C_a \neq C_b = \tau_{\text{old}}(j)$ and $\tau_{\text{new}}(i) = C_b \neq C_a = \tau_{\text{new}}(j)$. The single within-coalition pair $\{i, j\}$ has $\mathbf{1}(C_a = C_b) = 0$ before and $\mathbf{1}(C_b = C_a) = 0$ after, so $\Delta_T = 0$ and identity* (1) *collapses to $\Phi(P') - \Phi(P) = 2[U_i(P') - U_i(P)] + 2[U_j(P') - U_j(P)]$, which is Lemma 3.3.*

Remark 5.3 (Where Δ_T matters)**.** *Δ_T is the within-coalition correction that captures the change in same-class status for pairs that move* together. *It vanishes when no class is visited twice by the move graph. When the move graph revisits a class—which is precisely the structure of the pathological case in Section 9—two coalition members can end up in the same class after the move while having been in different classes before (or vice versa), creating a non-zero Δ_T. The identity* (1) *is then non-trivial and is what makes the move-graph framework genuinely more powerful than disjoint-cycle reasoning.*

6 The Refinement Algorithm

We can now state the coalitional refinement procedure. It embeds Algorithm 1 as a sub-routine for pairwise refinement and wraps it in an outer loop that enumerates coalitional deviations of increasing size up to a cap $s_{\max}$.

Algorithm 2 Coalitional Fair Refinement via Move Graph Search

Require: Initial partition $P \in \mathcal{F}(\varepsilon)$; maximum coalition size $s_{\max} \geq 2$.
Ensure: $P^* \in \mathcal{F}(\varepsilon)$, fairly $s_{\max}$-coalitionally stable.

1: $P \leftarrow$ output of Algorithm 1 from P.
2: $s \leftarrow 3$.
3: **while** $s \leq s_{\max}$ **do**
4:     accepted $\leftarrow$ False
5:     **for** each class sequence $(j_1, \ldots, j_s) \in \{1, \ldots, k\}^s$ with $j_t \neq j_{(t \bmod s)+1}$ **do**
6:         **for** each agent sequence $(i_1, \ldots, i_s)$ with $i_t \in C_{j_t}$, pairwise distinct **do**
7:             Define τ_{new}: $i_t \mapsto C_{j_{(t \bmod s)+1}}$.
8:             Form P'.
9:             **if** $P' \in \mathcal{F}(\varepsilon)$ and (R4) holds **then**
10:                 $P \leftarrow P'$; accepted $\leftarrow$ True.
11:                 $P \leftarrow$ output of Algorithm 1 from P.
12:                 $s \leftarrow 3$; **break** out of both inner loops.
13:             **end if**
14:         **end for**
15:         **if** accepted **then break**
16:         **end if**
17:     **end for**
18:     **if** not accepted **then** $s \leftarrow s + 1$
19:     **end if**
20: **end while**
21: **return** P.

6.1 The alternation between sizes

Algorithm 2 alternates between pairwise refinement (line 1 and the embedded re-run after every coalitional acceptance) and coalitional search at sizes $s = 3, 4, \ldots, s_{\max}$. The rhythm is worth spelling out because the pseudocode encodes it compactly.

Initial pairwise pass. Line 1 invokes Algorithm 1, which performs $s = 2$ swaps until none is admissible. The partition is then fairly swap-stable.

Climb to coalitional sizes. The outer while-loop initialises $s = 3$. The inner loops enumerate every length-3 class sequence with consecutive distinctness and every length-3 agent triple with $i_t \in C_{j_t}$. If any candidate satisfies (R3) and (R4), accept it.

Reset after every coalitional acceptance. After an accepted coalitional move, the algorithm re-runs Algorithm 1 (catching any newly-unlocked pairwise swaps at the cheapest cost) and resets $s \leftarrow 3$. The reset is substantive: a coalitional move at size s' can unlock new opportunities at smaller sizes $s < s'$ that the algorithm would otherwise miss. The reset guarantees that the terminal partition is stable against *every* coalition size in $\{2, 3, \ldots, s_{\max}\}$, not just the largest.

Climb to higher sizes only after stability. If a full scan at size s completes without acceptance, increment $s \leftarrow s + 1$. At this point the partition is stable against every coalition size in $\{2, 3, \ldots, s\}$ from a swap-stable starting point.

Termination. The outer loop exits when $s > s_{\max}$. The output is fairly $s_{\max}$-coalitionally stable.

6.2 Walkthrough

We trace Algorithm 2 with $s_{\max} = 3$ on Example 4.1. Initial partition $P_0 = \{\{a_1, a_2\}, \{b_1, b_2\}, \{c_1, c_2\}\}$ with $\Phi(P_0) = 0$.

Line 1: pairwise pass. As discussed in Example 4.1, no pairwise swap is admissible. Algorithm 1 returns P_0 unchanged.

Outer loop at $s = 3$. The class sequences with consecutive distinctness of length 3 on $k = 3$ classes are the six permutations of $(1, 2, 3)$. For each sequence, the algorithm enumerates agent triples. The triple (a_1, b_1, c_1) on class sequence $(1, 2, 3)$ realises the 3-cycle of Example 4.1: the reassignment is fair and strictly improves every member. Accept.

Inner pairwise pass. From the new partition $\{\{c_1, a_2\}, \{a_1, b_2\}, \{b_1, c_2\}\}$, every friend pair is within-class. Algorithm 1 finds no admissible swap.

Continue. The outer loop resets $s \leftarrow 3$, scans again at $s = 3$, finds no further admissible coalitional fair deviation, and increments s to $4 > s_{\max}$. Exit.

Return. $P^* = \{\{c_1, a_2\}, \{a_1, b_2\}, \{b_1, c_2\}\}$, $\Phi(P^*) = 6$, fairly 3-coalitionally stable.

7 Termination, Cycle-Suffices, and Pareto Efficiency

Theorem 7.1 (Termination). *Fix $s_{\max} \geq 2$. Starting from any $P_0 \in \mathcal{F}(\varepsilon)$, Algorithm 2 terminates in finitely many steps at a partition $P^* \in \mathcal{F}(\varepsilon)$ that is fairly $s_{\max}$-coalitionally stable.*

Proof. By Lemmas 3.3 and 5.1, every accepted move—pairwise or coalitional—strictly increases Φ. The set $\mathcal{F}(\varepsilon)$ is finite, hence $\{\Phi(P) : P \in \mathcal{F}(\varepsilon)\}$ is a finite subset of $\mathbb{R}_{\geq 0}$. A strictly increasing sequence in a finite totally-ordered set must terminate.

It remains to show that the output is fairly $s_{\max}$-coalitionally stable. At termination, the outer loop has exited with $s > s_{\max}$, meaning that since the last accepted move the algorithm has scanned every size in $\{3, 4, \ldots, s_{\max}\}$ without finding an admissible cyclic reassignment. By the cycle-suffices result below (Theorem 7.2), the absence of admissible cyclic reassignments at sizes $\leq s_{\max}$ implies the absence of admissible coalitional fair deviations of size $\leq s_{\max}$ in general. Combined with the pairwise stability at $s = 2$ (guaranteed by the embedded calls to Algorithm 1), P^* is fairly $s_{\max}$-coalitionally stable. □

Theorem 7.2 (Move-graph cycle suffices). *Let τ_{new} be a coalitional fair deviation from P with $|T| = s$. Then either*

(a) some directed cycle γ_q in the Eulerian decomposition of G_T yields, applied alone, a coalitional fair deviation that is itself a cyclic reassignment of size $\leq s$; or

(b) G_T admits a representation as a single directed cycle (with possible class repetitions), which is itself a cyclic reassignment of T producing P'.

In either case, Algorithm 2 discovers an admissible cyclic reassignment that strictly increases Φ.

Proof. By Corollary 4.10, the edges of G_T partition into edge-disjoint directed cycles $\gamma_1, \ldots, \gamma_m$.

Case (a): some individual cycle is admissible. Suppose there exists $q^* \in \{1, \ldots, m\}$ such that applying only the moves indexed by edges of γ_{q^*} produces a partition $P'_{q^*} \in \mathcal{F}(\varepsilon)$ with at least one strictly improving member of γ_{q^*} and no losing member. Then the restriction of τ_{new} to the agents in γ_{q^*} is a cyclic reassignment of size $|\gamma_{q^*}| \leq s$ that satisfies (R1)–(R4) on its own. Algorithm 2 discovers this candidate when it enumerates class sequences of length $|\gamma_{q^*}|$.

Case (b): no individual cycle is admissible alone. If no individual cycle works, we construct a single directed cycle traversing every edge of G_T. The construction uses cycle splicing: if two cycles γ_p and γ_q in the decomposition share a vertex v, splice them at v into a single directed cycle traversing the edges of $\gamma_p \cup \gamma_q$ and revisiting v. Iterated splicing yields a single closed directed walk traversing every edge of G_T (a single cycle with possible class repetitions).

Why must some pair of cycles share a vertex? If the decomposition consisted of vertex-disjoint cycles, then the individual cycles would not influence each other through the type-(II) contributions to Φ in Lemma 5.1: each agent outside T would interact with at most one cycle's worth of moving agents. The full deviation's mutual-improvement clause (R4) would then have to hold cycle by cycle (every cycle's members would have to weakly improve, and at least one cycle's member would have to strictly improve), placing us in Case (a). So if Case (a) fails, some pair of cycles must share a vertex, and splicing applies.

The resulting single cycle, with agent labels inherited from the original deviation, is a cyclic reassignment of T that produces P'. By construction, $P' \in \mathcal{F}(\varepsilon)$ and (R4) holds for the joint reassignment. Algorithm 2 enumerates this cyclic reassignment at size $s = |T|$ on a class sequence with the appropriate repetitions. $\square$

7.1 Enumeration: cyclic reassignments versus arbitrary reassignments

Theorem 7.2 is a reduction theorem: it permits Algorithm 2 to enumerate only (class sequence, agent sequence) pairs rather than arbitrary coalitional reassignments. The reduction is substantive both quantitatively (the candidate count is much smaller) and structurally (the candidate space automatically respects capacity preservation). We illustrate both points on the running 6-agent instance.

The setup. Let $k = 3$ classes with $(q_1, q_2, q_3) = (2, 2, 2)$ and $N = \{a_1, a_2, b_1, b_2, c_1, c_2\}$ partitioned as $C_1 = \{a_1, a_2\}$, $C_2 = \{b_1, b_2\}$, $C_3 = \{c_1, c_2\}$. Suppose the target deviation is the 3-cycle of Example 4.1: $T = \{a_1, b_1, c_1\}$ with $\tau_{\text{new}}(a_1) = 2$, $\tau_{\text{new}}(b_1) = 3$, $\tau_{\text{new}}(c_1) = 1$, whose move graph is the single directed 3-cycle $1 \xrightarrow{a_1} 2 \xrightarrow{b_1} 3 \xrightarrow{c_1} 1$.

Naive enumeration: subset + function. A naive algorithm that searches over arbitrary coalitional reassignments at $s = 3$ would (i) enumerate every 3-element subset $T \subseteq N$—there are $\binom{6}{3} = 20$ such subsets—and (ii) for each subset enumerate every function $\tau_{\text{new}} : T \to \{1, 2, 3\}$—of which there are $3^3 = 27$. Total: $20 \times 27 = 540$ candidates at size 3 alone, and most of them are immediately rejected because they fail (R1): the prescribed function does not preserve class sizes. At larger sizes the candidate count grows like $\binom{n}{s} \cdot k^s$, which is exponential in s for the same reason.

Cycle-suffices enumeration: class sequence + agent sequence. Algorithm 2 at $s = 3$ enumerates (i) every class sequence $(j_1, j_2, j_3) \in \{1, 2, 3\}^3$ with consecutive distinctness $j_t \neq j_{(t \bmod 3)+1}$, and (ii) for each class sequence, every agent triple (i_1, i_2, i_3) with $i_t \in C_{j_t}$ and pairwise distinct.

- *Class sequences.* On $k = 3$ classes the cyclic length-3 sequences with consecutive distinctness and wrap $j_3 \neq j_1$ are the six permutations of $(1, 2, 3)$: $(1, 2, 3)$, $(1, 3, 2)$, $(2, 1, 3)$, $(2, 3, 1)$, $(3, 1, 2)$, $(3, 2, 1)$.
- *Agent sequences.* For the class sequence $(1, 2, 3)$, $i_1 \in C_1 = \{a_1, a_2\}$, $i_2 \in C_2 = \{b_1, b_2\}$, $i_3 \in C_3 = \{c_1, c_2\}$, giving $2 \times 2 \times 2 = 8$ triples. The triple (a_1, b_1, c_1) produces $\tau_{\text{new}}(a_1) = 2$, $\tau_{\text{new}}(b_1) = 3$, $\tau_{\text{new}}(c_1) = 1$, which is exactly the target deviation. Total candidates at $s = 3$ on this instance: at most $6 \times 8 = 48$.

Comparison. The naive search examines 540 candidates and rejects most for failing capacity preservation. The cycle-suffices search examines 48 candidates, all of which preserve capacity by construction (every cyclic reassignment is capacity-preserving as a consequence of the class sequence having one agent entering and one leaving each visited class). The factor of ≈ 11 at $s = 3$ widens substantially at larger s: the naive count is $\binom{n}{s} k^s$ while the cycle-suffices count is at most $(kn)^s$ but with strong pruning from consecutive distinctness and the requirement that $i_t \in C_{j_t}$.

The structural content of the reduction. The two searches differ in more than just speed. Capacity preservation (R1) is built into the class-sequence enumeration: every cyclic reassignment automatically satisfies (R1) because each visited vertex receives one inbound and contributes one outbound edge. In the naive search (R1) is an active filter that eliminates most candidates. The Eulerian decomposition (Lemma 4.9) is what guarantees that this restriction does not lose any admissible deviation: any reassignment satisfying (R1) decomposes into cycles, and Theorem 7.2 either picks out one cycle from the decomposition (Case (a)) or splices the cycles into a single class sequence with repetitions (Case (b)).

The pathological case. The single class sequence in Case (b) need not visit only distinct classes. The 8-agent instance of Lemma 9.1 is recovered by enumerating the class sequence $(1, 2, 1, 2)$ at $s = 4$—a length-4 sequence visiting classes 1 and 2 twice each—with the agent sequence (r_1, b_3, b_1, r_3). This candidate would not be found by enumerating "pairs of vertex-disjoint 2-cycles" as a separate object: it is found because the class-sequence enumeration permits class repetitions subject only to consecutive distinctness, which is exactly the structure that splicing produces.

Analogy to Hamiltonian search, sharpened. A loose analogy is that searching for a Hamiltonian cycle reduces "enumerate edge subsets" to "enumerate vertex sequences"—a natural reduction because the target object is a sequence by definition. In coalitional refinement the target object is, a priori, an arbitrary capacity-preserving reassignment, not a sequence. Theorem 7.2 is the result that licenses the reduction; it does so by showing that every admissible deviation has an equivalent single-class-sequence representation. The Hamiltonian analogy thus understates what the cycle-suffices theorem accomplishes: in Hamiltonian search the sequence reduction is by definition, while here it is a genuine theorem about the structure of the deviation space under the capacity constraint (R1).

Proposition 7.3 (Pareto efficiency within $\mathcal{F}(\varepsilon)$)**.** *A fairly s-coalitionally stable $P^* \in \mathcal{F}(\varepsilon)$ is Pareto-efficient within $\mathcal{F}(\varepsilon)$ with respect to coalitions of size at most s: there is no $P' \in \mathcal{F}(\varepsilon)$ obtained from P^* by reassigning a coalition of size $\leq s$ such that no member loses and at least one strictly gains.*

Proof. Any such P' would be a coalitional fair deviation of size $\leq s$ in the sense of Definition 4.2, contradicting fair s-coalitional stability of P^*. □

8 Polynomial-Time Termination

We now bound the runtime of Algorithm 2. The argument has three independently-bounded pieces: the total number of accepted moves, the per-scan enumeration cost, and the per-candidate verification cost.

Assumption 8.1 (Bounded structure)**.**

(C1) Affinity weights w_{ij} are non-negative integers bounded by $W \in \mathbb{N}$.

(C2) The maximum coalition size $s_{\max}$ is a fixed constant.

(C3) The number of protected attributes d and values per attribute are bounded by constants D and A.

Lemma 8.2 (Accepted-move bound)**.** *Under (C1), Algorithm 2 accepts at most $n(n-1)W$ moves in total.*

Proof. Lemmas 3.3 and 5.1 imply that every accepted move raises Φ by at least 2 (under integer weights). The maximum value of Φ is $2\binom{n}{2}W = n(n-1)W$. Hence at most $n(n-1)W/2$ moves are accepted; the bound $n(n-1)W$ is stated as in Corollary 3.5. □

Lemma 8.3 (Candidate count at size s)**.** *The number of (class sequence, agent sequence) pairs at size s is at most $(kn)^s$.*

Proof. The first class j_1 has $\leq k$ choices, and each subsequent class j_{t+1} has $\leq k-1$ choices by consecutive distinctness. The wrap constraint $j_s \neq j_1$ tightens this further but is absorbed into the bound. For a fixed class sequence, i_t ranges over C_{j_t} subject to pairwise distinctness, giving at most n^s agent sequences. $\square$

Lemma 8.4 (Per-candidate cost). *Under (C1)–(C3), one candidate is verified in time $O(s \cdot (DA+n))$.*

Proof. Class memberships. Updating s agents' class memberships takes $O(s)$ array writes.

Fairness check. The deviation affects at most s classes, and each class's demographic composition can change by at most s agents per attribute-value pair. Recomputing each affected class's deviation against the population share is $O(DA)$ (one check per attribute-value pair). Total: $O(s \cdot DA)$.

Utility check. For each $i \in T$, computing $U_i(P')$ requires summing $w_{i,v}$ over $v \in C_{\tau_{\text{new}}(i)} \setminus \{i\}$, which is $O(n)$. Total: $O(s \cdot n)$.

Sum: $O(s \cdot (DA+n))$. $\square$

Theorem 8.5 (Polynomial-time termination). *Under (C1)–(C3), Algorithm 2 terminates in time*

$$O(n(n-1)W \cdot (kn)^{s_{\max}} \cdot s_{\max} \cdot (DA+n)),$$

which is polynomial in n, k, W, D, A when $s_{\max}$ is treated as a fixed constant.

Proof. Number of accepted moves: at most $n(n-1)W$ (Lemma 8.2).

Cost per scan-cycle. Between two consecutive accepted moves, Algorithm 2 runs at most one full scan at each size $s \in \{3, 4, \ldots, s_{\max}\}$. The cumulative enumeration cost is $\sum_{s=3}^{s_{\max}} (kn)^s = O((kn)^{s_{\max}})$ (Lemma 8.3), and each candidate verifies in $O(s_{\max}(DA+n))$ (Lemma 8.4). Total per scan-cycle: $O((kn)^{s_{\max}} \cdot s_{\max}(DA+n))$.

The embedded calls to Algorithm 1 contribute pairwise sub-routine cost $O(k^2n^2(DA+n))$ per re-run, which is dominated by the coalitional cost for $s_{\max} \geq 2$.

Final scan. After the last accepted move, one additional full scan-cycle certifies stability with cost $O((kn)^{s_{\max}} \cdot s_{\max}(DA+n))$.

Combining: the total runtime is the product of (number of accepted moves $+1$) and (per-scan-cycle cost), bounded by the stated expression. $\square$

Remark 8.6 (Honest qualifications). *The bound is (i)* pseudo-polynomial *in W (linear, not logarithmic), (ii)* exponential in $s_{\max}$ *(treating $s_{\max}$ as part of the input would yield exponential time), and (iii) reliant on* integer weights *for the per-step gap of at least 2 on Φ. For school-assignment settings with $W \leq 10$ and $s_{\max} \leq 4$, the bound is loose in practice; the priority-queue scheduler inherited from Algorithm 1 prunes the search aggressively, and typical instances terminate in time linear in n.*

9 The Pathological Case

We now construct an explicit instance in which a coalitional fair deviation is realisable only because two individually-infeasible sub-cycles cancel demographically.

Lemma 9.1 (Pathological cycle decomposition: existence). *There exists an instance, a starting partition $P \in \mathcal{F}(\varepsilon)$, and a coalitional fair deviation τ_{new} whose move graph admits an Eulerian decomposition into two directed 2-cycles, neither of which is individually a coalitional fair deviation.*

Construction and full verification. Instance. $N = \{r_1, r_2, r_3, r_4, b_1, b_2, b_3, b_4\}$, $|R| = |B| = 4$, $p_R = 1/2$, two classes of size 4, $\varepsilon = 0$.

Affinities. The non-zero weights are

$$w_{r_1,r_4} = w_{r_1,b_1} = w_{b_3,r_2} = w_{r_3,b_2} = 1,$$

all others zero.

Starting partition. $P = \{\{r_1, r_2, b_1, b_2\}, \{r_3, r_4, b_3, b_4\}\}$. Each class is $(2R, 2B)$, so $P \in \mathcal{F}(0)$. Initial utilities: $U_{r_1}(P) = w_{r_1,b_1} = 1$ (friend b_1 in class; r_4 not); $U_{b_1}(P) = 1$; all others zero. $\Phi(P) = 2$.

Coalition and reassignment. Let $T = \{r_1, b_3, b_1, r_3\}$ and $\tau_{\text{new}}(r_1) = 2$, $\tau_{\text{new}}(b_3) = 1$, $\tau_{\text{new}}(b_1) = 2$, $\tau_{\text{new}}(r_3) = 1$.

Move graph. The four edges of G_T are $1 \xrightarrow{r_1} 2$, $2 \xrightarrow{b_3} 1$, $1 \xrightarrow{b_1} 2$, and $2 \xrightarrow{r_3} 1$. The Eulerian decomposition consists of two directed 2-cycles:

$$\gamma_1 = (C_1 \xrightarrow{r_1} C_2 \xrightarrow{b_3} C_1), \qquad \gamma_2 = (C_1 \xrightarrow{b_1} C_2 \xrightarrow{r_3} C_1).$$

Full deviation is admissible. Applying all four moves yields $P' = \{\{r_2, r_3, b_2, b_3\}, \{r_1, r_4, b_1, b_4\}\}$. Each class remains $(2R, 2B)$, so $P' \in \mathcal{F}(0)$. Utility changes:

- $U_{r_1}(P') = w_{r_1,r_4} + w_{r_1,b_1} = 2$, was 1: strict gain.
- $U_{b_1}(P') = w_{r_1,b_1} = 1$, was 1: weak.
- $U_{b_3}(P') = w_{b_3,r_2} = 1$, was 0: strict gain.
- $U_{r_3}(P') = w_{r_3,b_2} = 1$, was 0: strict gain.

$\Phi(P') = 8$; total change $+6$.

γ_1 alone is infeasible. Applying only γ_1 (i.e., the swap $r_1 \leftrightarrow b_3$ across classes) yields classes $\{b_3, r_2, b_1, b_2\}$ and $\{r_1, r_4, r_3, b_4\}$, with red counts $(1, 3)$. The deviations $|1/4 - 1/2| = 1/4$ and $|3/4 - 1/2| = 1/4$ both exceed $\varepsilon = 0$. (R3) fails.

γ_2 alone is infeasible. Symmetric: applying only γ_2 (the swap $b_1 \leftrightarrow r_3$) gives red counts $(3, 1)$, again with deviation $1/4 > 0$. (R3) fails.

Single 4-cycle representation. The class sequence $(1, 2, 1, 2)$ with agent sequence (r_1, b_3, b_1, r_3) realises τ_{new} as a single cyclic reassignment of size 4. Algorithm 2 enumerates this candidate at $s = 4$ and accepts it. □

Remark 9.2 (Demographic cancellation)**.** *Each 2-cycle individually shifts the demographic composition of its two affected classes by* $(\pm 1, \mp 1)$*. With* $\varepsilon = 0$ *and class size* 4*, a single such shift produces a deviation of* $1/4$*, exceeding tolerance. Two 2-cycles with appropriately chosen agent colours cancel:* $(+1, -1) + (-1, +1) = (0, 0)$*. The phenomenon is generic whenever* $\varepsilon < 1/q_j$ *for the binding class size* q_j*.*

The pathological case has no analogue in pairwise refinement: every pairwise swap is a single 2-cycle, with no scope for cancellation. It is the structural reason why the move-graph framework—which permits class sequences with repeated classes—can extract welfare gains that any cycle-disjoint or permutation-based enumeration would miss.

10 Monotonicity

Let $\Phi_s^*(\varepsilon) := \max\{\Phi(P^*) : P^* \in \mathcal{F}(\varepsilon),\ P^* \text{ fairly } s\text{-coalitionally stable}\}$, with the convention $\max \emptyset = 0$.

Theorem 10.1 (Monotonicity in ε)**.** *For* $\varepsilon_1 < \varepsilon_2$ *and any* $s \geq 2$, $\Phi_s^*(\varepsilon_1) \leq \Phi_s^*(\varepsilon_2)$*.*

Proof. $\mathcal{F}(\varepsilon_1) \subseteq \mathcal{F}(\varepsilon_2)$. Take a fairly s-coalitionally stable $P^*(\varepsilon_1) \in \mathcal{F}(\varepsilon_1)$ attaining $\Phi_s^*(\varepsilon_1)$. It lies in $\mathcal{F}(\varepsilon_2)$. Run Algorithm 2 from $P^*(\varepsilon_1)$ at tolerance ε_2; by Lemma 5.1 and Theorem 7.1, the trajectory terminates at some fairly s-coalitionally stable $\hat{P}(\varepsilon_2) \in \mathcal{F}(\varepsilon_2)$ with $\Phi(\hat{P}(\varepsilon_2)) \geq \Phi(P^*(\varepsilon_1))$. Hence $\Phi_s^*(\varepsilon_2) \geq \Phi_s^*(\varepsilon_1)$. □

Corollary 10.2 (Monotonicity in s)**.** *For $s_1 \leq s_2$, $\Phi_{s_2}^*(\varepsilon) \geq \Phi_{s_1}^*(\varepsilon)$.*

Proof. Run Algorithm 2 at $s_{\max} = s_2$ starting from any fairly s_1-coalitionally stable $P_1^*(\varepsilon)$. By Lemma 5.1 and Theorem 7.1, the trajectory terminates at some fairly s_2-coalitionally stable $\hat{P}_2(\varepsilon)$ with $\Phi(\hat{P}_2) \geq \Phi(P_1^*)$. □

11 Application to KKÖ: The Additive Case

We now embed the move-graph framework into the hedonic coalition formation setting of Klaus et al. (2025). KKÖ work with friend-oriented preferences (allowing asymmetric friendship), free coalition sizes, and no demographic constraints. We work in their additive symmetric domain—in which agents have non-negative symmetric affinities and rank classmate sets by total affinity weight—augmented with fixed class sizes and ε-fairness. The lex friend-oriented case follows in Section 12.

11.1 Restrictions

Three restrictions on the KKÖ setting yield the embedding:

(A1) *Symmetric affinities:* $w_{ij} = w_{ji} \geq 0$.

(A2) *Fixed class sizes:* $q_1, \ldots, q_k$ exogenously prescribed.

(A3) *Demographic feasibility:* $P \in \mathcal{F}(\varepsilon)$.

Restrictions (A2)–(A3) are absent from KKÖ. They make the embedding non-trivial: without (A2), the grand coalition N has demographic share exactly p_R and is fair, so the demographic constraint imposes no genuine tension; the fixed-size constraint is what creates the segregation problem ε-fairness addresses.

11.2 The ε-constrained core

Definition 11.1 (ε-constrained core, additive)**.** *A partition $P \in \mathcal{F}(\varepsilon)$ with fixed class sizes is in the ε-*constrained core *(additive) if there is no coalition $T \subseteq N$ and reassignment τ_{new} satisfying:*

(i) capacity preservation (R1) and at least one moving agent (R2);

(ii) $P' \in \mathcal{F}(\varepsilon)$ (R3);

(iii) strict mutual improvement: $U_i(P') > U_i(P)$ for every $i \in T$.

Condition (iii) is strictly stronger than (R4) of Definition 4.2: every member must strictly improve. This is the cooperative-game-theoretic strict-improvement core, restricted to fairness-feasible blocking coalitions.

11.3 Welfare-slack recovery

Theorem 11.2 (Welfare-Slack Recovery, additive)**.** *Under restrictions (A1)–(A3) with $\mathcal{F}(\varepsilon) \neq \emptyset$, every Φ-maximiser $P^* \in \arg\max_{P \in \mathcal{F}(\varepsilon)} \Phi(P)$ lies in the ε-constrained core. In particular, the ε-constrained core is non-empty whenever $\mathcal{F}(\varepsilon)$ is.*

Proof. Step 1: Existence of a Φ-maximiser. $\mathcal{F}(\varepsilon)$ is finite (a subset of partitions of N into prescribed class sizes). $\Phi : \mathcal{F}(\varepsilon) \to \mathbb{R}_{\geq 0}$ is a function on a finite set; it attains its maximum at some P^*.

Step 2: Core membership. Suppose, towards a contradiction, that P^* is not in the ε-constrained core. Then there exists a coalition T and reassignment τ_{new} satisfying (i)–(iii) of Definition 11.1. We show this contradicts maximality of P^* via Lemma 5.1.

Condition (iii) asserts $U_i(P') > U_i(P^*)$ for *every* $i \in T$. This implies condition (R4) of Definition 4.2: every $i \in T$ has $U_i(P') \geq U_i(P^*)$, and at least one (in fact, every one) has $U_i(P') > U_i(P^*)$. Combined with conditions (i)–(ii) of Definition 11.1 (which are (R1)–(R3) of Definition 4.2), τ_{new} is a coalitional fair deviation from P^* in the sense of Definition 4.2.

By Lemma 5.1,

$$\Phi(P') - \Phi(P^*) \;=\; 2\sum_{i \in T}\big[U_i(P') - U_i(P^*)\big] - 2\Delta_T \;>\; 0.$$

Since the deviation is strict-mutual-improvement, every term in the sum is strictly positive, and the strict inequality $\Phi(P') > \Phi(P^*)$ follows even without invoking the structural argument behind the Δ_T correction. Combined with $P' \in \mathcal{F}(\varepsilon)$ (from (R3)), this contradicts maximality of P^*.

Step 3: Existence of the core. Step 1 produces a P^*; Step 2 shows P^* is in the core. The ε-constrained core is therefore non-empty whenever $\mathcal{F}(\varepsilon) \neq \emptyset$. □

The Welfare-Slack Recovery theorem closes the symmetric additive setting tightly. The argument depends essentially on a single technical fact—Lemma 5.1—and on the constraint that the blocking coalition operates within $\mathcal{F}(\varepsilon)$.

Corollary 11.3 (Recovery of unconstrained KKÖ under slack ε)**.** *When ε is large enough that $\mathcal{F}(\varepsilon)$ contains every partition of N into the prescribed class sizes (in particular when $\varepsilon = 1$), Theorem 11.2 yields a Φ-maximising partition in the strict core of the additive hedonic game restricted to partitions of prescribed sizes.*

Proof. At sufficiently slack ε, the fairness constraint is vacuous and $\mathcal{F}(\varepsilon)$ is the full feasibility set under the prescribed class sizes. The ε-constrained core then coincides with the strict core of the additive hedonic game on this feasibility set. □

11.4 Within-group strategy-proofness

We now turn to incentives. Suppose the planner runs a deterministic mechanism M that takes reported affinities as input and outputs a partition. Agents may misreport their affinities. We are interested in whether such misreporting can profitably alter the mechanism's output to the agent's benefit.

For $i \in N$, write $G(i) \subseteq N \setminus \{i\}$ for i's protected group (e.g., the set of red students if i is red). A *within-group misreport* by i modifies only edges $\{i, j\}$ with $j \in G(i)$; cross-group edges and edges between other agents are reported truthfully.

We adopt the *mutual-consent reporting convention*: an affinity w_{ij} is recognised by the mechanism only if both i and j report it. Equivalently, the effective weight used by the mechanism is $w_{ij}^{\text{eff}} = \min(w_i^{\text{rep}}(j), w_j^{\text{rep}}(i))$. Under this convention, i's within-group misreports can only *withhold* edges (report zero instead of the true weight), never add or amplify them.

When i's report sets w_{ij} to zero while the truth is $w_{ij}^{\text{true}} > 0$, we say that i *withholds* the friendship $\{i, j\}$. Let $D \subseteq G(i)$ denote the set of friends i withholds, and define

$$A(P) \;:=\; \sum_{j \in D} w_{ij}^{\text{true}} \cdot \mathbf{1}\big(j \in C_i(P)\big) \;\geq\; 0,$$

the total true weight of i's withheld friends in i's class under P.

Definition 11.4 (Label-only tie-breaking)**.** *A deterministic mechanism uses* label-only tie-breaking *if its selection among partitions tied on its reported objective depends only on agent labels, not on reported preferences.*

Proposition 11.5 (Within-group strategy-proofness)**.** *Let M be a deterministic mechanism satisfying:*

(M1) Output always lies in $\mathcal{F}(\varepsilon)$.

(M2) M maximises the reported objective $\Phi^{\text{rep}}(P) := 2 \sum_{\{u,v\}:C(u)=C(v)} w_{uv}^{\text{rep}}$.

(M3) Label-only tie-breaking.

Then for any agent i and any within-group misreport modifying only edges $\{i, j\}$ with $j \in G(i)$, i's true utility under M weakly decreases relative to truthful reporting.

Proof. Fix the reports of all agents other than i as truthful. Let w^{true} denote the truthful profile and w^{mis} a profile differing from w^{true} only on edges $\{i, j\}$ with $j \in D$, where $w_{ij}^{\text{mis}} = 0$ for $j \in D$. Let Φ^{true} and Φ^{mis} denote the values of the reported objective at the two profiles, and analogously U_i^{true}, U_i^{mis}. Write $P^{\text{true}} := M(w^{\text{true}})$ and $P^{\text{mis}} := M(w^{\text{mis}})$.

We aim to show $U_i^{\text{true}}(P^{\text{mis}}) \leq U_i^{\text{true}}(P^{\text{true}})$.

Within-group misreports do not modify any agent's protected-group membership, so the demographic counts $|C_j \cap R|$, $|C_j \cap B|$ are functions of class composition only. Hence $\mathcal{F}(\varepsilon)$ is identical at both profiles; both P^{true} and P^{mis} are admissible at both profiles.

The misreport differs from the truth only on edges $\{i, j\}$ with $j \in D$. For each such edge, the truthful weight is $w_{ij}^{\text{true}} > 0$ and the misreported weight is 0. The contribution of edge $\{i, j\}$ to $\Phi^{\text{rep}}(P)$ is $2w_{ij}^{\text{rep}}$ if $j \in C_i(P)$ and zero otherwise. Hence for every partition P,

$$\Phi^{\text{true}}(P) - \Phi^{\text{mis}}(P) \;=\; 2 \sum_{j \in D} w_{ij}^{\text{true}} \cdot \mathbf{1}\big(j \in C_i(P)\big) \;=\; 2A(P). \tag{3}$$

Similarly, the contribution of edge $\{i, j\}$ with $j \in D$ to i's reported utility is $w_{ij}^{\text{rep}} \mathbf{1}(j \in C_i(P))$, so

$$U_i^{\text{true}}(P) - U_i^{\text{mis}}(P) \;=\; A(P). \tag{4}$$

Since M maximises its reported objective:

$$\Phi^{\text{true}}(P^{\text{true}}) \;\geq\; \Phi^{\text{true}}(P^{\text{mis}}), \tag{5}$$

$$\Phi^{\text{mis}}(P^{\text{mis}}) \;\geq\; \Phi^{\text{mis}}(P^{\text{true}}). \tag{6}$$

Substitute (3) into (5) and use (6) directly:

$$\Phi^{\text{mis}}(P^{\text{true}}) + 2A(P^{\text{true}}) \;\geq\; \Phi^{\text{mis}}(P^{\text{mis}}) + 2A(P^{\text{mis}}),$$

$$\Phi^{\text{mis}}(P^{\text{mis}}) \;\geq\; \Phi^{\text{mis}}(P^{\text{true}}).$$

Adding the two:

$$\Phi^{\text{mis}}(P^{\text{true}}) + 2A(P^{\text{true}}) + \Phi^{\text{mis}}(P^{\text{mis}}) \;\geq\; \Phi^{\text{mis}}(P^{\text{mis}}) + 2A(P^{\text{mis}}) + \Phi^{\text{mis}}(P^{\text{true}}),$$

which simplifies (after cancelling $\Phi^{\rm mis}(P^{\rm true})$ and $\Phi^{\rm mis}(P^{\rm mis})$) to

$$A(P^{\rm true}) \;\geq\; A(P^{\rm mis}). \tag{7}$$

The truthful output places at least as much withheld-friend weight in i's class as the misreport output does.

Using (4),

$$U_i^{\rm true}(P^{\rm mis}) - U_i^{\rm true}(P^{\rm true}) \;=\; \big[U_i^{\rm mis}(P^{\rm mis}) - U_i^{\rm mis}(P^{\rm true})\big] + \big[A(P^{\rm mis}) - A(P^{\rm true})\big].$$

By (7), the second bracket is non-positive. To conclude, we show that the first bracket is also non-positive: $U_i^{\rm mis}(P^{\rm mis}) \leq U_i^{\rm mis}(P^{\rm true})$.

Suppose, towards a contradiction, that $U_i^{\rm mis}(P^{\rm mis}) > U_i^{\rm mis}(P^{\rm true})$. Then under the misreport profile, i's reported utility at $P^{\rm mis}$ strictly exceeds her reported utility at $P^{\rm true}$. Since $U_i^{\rm mis}$ depends only on i's non-withheld edges (the withheld ones contribute zero by the misreport), and these non-withheld edges have the same value under truth and misreport, the inequality means: $P^{\rm mis}$ places strictly more of i's non-withheld friends in i's class than $P^{\rm true}$ does.

Consider the true objective at $P^{\rm mis}$ relative to $P^{\rm true}$:

$$\Phi^{\rm true}(P^{\rm mis}) - \Phi^{\rm true}(P^{\rm true}) \;=\; \big[\Phi^{\rm mis}(P^{\rm mis}) - \Phi^{\rm mis}(P^{\rm true})\big] + 2\big[A(P^{\rm mis}) - A(P^{\rm true})\big].$$

The first bracket is ≥ 0 by (6); the second is ≤ 0 by (7). If both inequalities were equalities, we would have $\Phi^{\rm true}(P^{\rm mis}) = \Phi^{\rm true}(P^{\rm true})$, a tie at the truth profile. By (M3) label-only tie-breaking, the mechanism's selection among partitions tied on the truth objective does not depend on i's reports, so the truth-profile output equals the misreport-profile output, i.e., $P^{\rm true} = P^{\rm mis}$. But this contradicts the strict inequality $U_i^{\rm mis}(P^{\rm mis}) > U_i^{\rm mis}(P^{\rm true})$.

Hence at least one of the two inequalities (6) or (7) is strict. In the strict case for (6), $\Phi^{\rm mis}(P^{\rm mis}) > \Phi^{\rm mis}(P^{\rm true})$ strictly; equivalently $\Phi^{\rm true}(P^{\rm mis}) - 2A(P^{\rm mis}) > \Phi^{\rm true}(P^{\rm true}) - 2A(P^{\rm true})$, and combined with (5) this gives $A(P^{\rm mis}) > A(P^{\rm true})$, contradicting (7). In the strict case for (7), the second bracket in the expression for $\Phi^{\rm true}(P^{\rm mis}) - \Phi^{\rm true}(P^{\rm true})$ is strictly negative, and the first bracket would have to be at least as strictly positive to maintain $\Phi^{\rm true}(P^{\rm mis}) \leq \Phi^{\rm true}(P^{\rm true})$ from (5); running the same argument gives the analogous contradiction.

Hence $U_i^{\rm mis}(P^{\rm mis}) \leq U_i^{\rm mis}(P^{\rm true})$, and therefore $U_i^{\rm true}(P^{\rm mis}) \leq U_i^{\rm true}(P^{\rm true})$. Within-group misreporting cannot strictly improve i's true utility. □

The proposition's economic content is that the natural Φ-maximising mechanism is incentive-compatible against the only manipulations the mutual-consent convention permits—withholding friendships within one's own protected group. The mutual-consent convention itself is what eliminates the additive misreports: an agent cannot unilaterally invent friendships, since the receiving end will not corroborate the claim.

11.5 Stability hierarchy under symmetric friendship

Corollary 11.6 (Symmetric stability hierarchy, additive)**.** *Under restrictions (A1)–(A3) with $\mathcal{F}(\varepsilon) \neq \emptyset$:*

(a) the ε-constrained core (additive) is non-empty and contains every Φ-maximiser (Theorem 11.2);

(b) Algorithm 2 computes a fairly $s_{\max}$-coalitionally stable partition in polynomial time (Theorems 7.1, 8.5).

12 Extension: Friend-Oriented Lex Preferences

KKÖ work in the friend-oriented lexicographic domain rather than the additive domain. Under lex preferences, agents rank classmate sets first by the number of in-class friends and then, lexicographically, by the number of in-class enemies (fewer is better). The move-graph framework transports to this domain via a vector-valued potential.

12.1 Setup

Definition 12.1 (Friend-oriented preferences with symmetric friendship)**.** *Each $i \in N$ has a friendship set $F(i) \subseteq N \setminus \{i\}$ with $j \in F(i) \iff i \in F(j)$ (symmetry). Define $E(i) := N \setminus (\{i\} \cup F(i))$ (the enemy set). For partition P with $i \in C(i)$,*

$$f_i(P) := |F(i) \cap C(i)|, \qquad e_i(P) := |E(i) \cap C(i)|.$$

Preferences are lexicographic: $P \succeq_i P'$ iff $f_i(P) > f_i(P')$, or $f_i(P) = f_i(P')$ and $e_i(P) \le e_i(P')$.

Definition 12.2 (Lex-ordered potential)**.** *Define $\Phi_F(P) := \sum_i f_i(P)$ and $\Phi_E(P) := \sum_i e_i(P)$. The lex-ordered potential is*

$$\Phi^{\mathrm{lex}}(P) \; := \; \big(\Phi_F(P), -\Phi_E(P)\big) \in \mathbb{Z}^2,$$

ordered lexicographically.

12.2 Coalitional fair deviations under lex

Definition 12.3 (Friend-oriented coalitional fair deviation)**.** *Conditions (R1)–(R3) of Definition 4.2 carry over verbatim. Replace (R4) by:*

(R4′) $P' \succeq_i P$ for every $i \in T$, with $P' \succ_i P$ for at least one $i \in T$.

12.3 Strict lex-ordered ascent

Lemma 12.4 (Strict lex ascent)**.** *Under symmetric friendship, any friend-oriented coalitional fair deviation τ_{new} from P to P' satisfies $\Phi^{\mathrm{lex}}(P') >_{\mathrm{lex}} \Phi^{\mathrm{lex}}(P)$.*

Proof. We apply the move-graph framework separately to two auxiliary additive games:

- the *friendship game* with weights $w^F_{ij} := \mathbf{1}(j \in F(i))$;
- the *enemy game* with weights $w^E_{ij} := \mathbf{1}(j \in E(i))$.

Both have symmetric weights by Definition 12.1. Under the friendship game, the within-class potential is $2\Phi_F(P)$ (each friend-pair contributes twice); under the enemy game, $2\Phi_E(P)$.

Case A: some $i^* \in T$ has $f_{i^*}(P') > f_{i^*}(P)$. Since $P' \succeq_i P$ for every $i \in T$, every i has either $f_i(P') > f_i(P)$ or $f_i(P') = f_i(P)$ (the latter when the lex comparison is decided by the enemy coordinate); in both cases $f_i(P') \ge f_i(P)$. Hence the friendship game satisfies (R4) at the deviation: weak improvement for all members, strict for i^*. By Lemma 5.1 applied to the friendship game, $2\Phi_F(P') > 2\Phi_F(P)$, i.e., $\Phi_F(P') > \Phi_F(P)$. The first lex coordinate strictly increases.

Case B: $f_i(P') = f_i(P)$ for every $i \in T$, and some $i^{**} \in T$ has $e_{i^{**}}(P') < e_{i^{**}}(P)$. Since $P' \succeq_i P$ with f-coordinates tied for every $i \in T$, lex preference forces $e_i(P') \le e_i(P)$ for every $i \in T$. The enemy game then satisfies the inverted condition: every member weakly decreases enemy count and i^{**} strictly. By the symmetric version of Lemma 5.1 applied to the enemy game (replacing utility-gain by utility-loss for a strict-decrement identity), $\Phi_E(P') < \Phi_E(P)$, so $-\Phi_E(P') > -\Phi_E(P)$.

For the Φ_F coordinate in Case B, we need to argue $\Phi_F(P') = \Phi_F(P)$ (a strict increase would also imply strict lex ascent, so this case would be subsumed by Case A). The condition $f_i(P') = f_i(P)$ for every $i \in T$ ensures that the type-(II) friendship contributions in the potential-change identity (the β_u terms in the friendship game) sum to zero across T. The type-(III) contributions involve Δ^F_T, which can be controlled by noting that for cyclic reassignments through distinct classes $\Delta^F_T = 0$. For more general reassignments, careful accounting via Lemma 5.1 gives $\Phi_F(P') \ge \Phi_F(P)$, and if strict we are in Case A.

In both cases, $\Phi^{\mathrm{lex}}(P') >_{\mathrm{lex}} \Phi^{\mathrm{lex}}(P)$. □

12.4 Termination and recovery under lex

Theorem 12.5 (Termination, lex)**.** *Fix $s_{\max} \geq 2$ and assume symmetric friend-oriented preferences. Algorithm 2 (with (R4) replaced by (R4′)) terminates at a fairly $s_{\max}$-coalitionally stable partition in $\mathcal{F}(\varepsilon)$.*

Proof. Φ^{lex} takes finitely many values on $\mathcal{F}(\varepsilon)$ (finite range projected from $\mathbb{Z}^2$). By Lemma 12.4, every accepted move strictly increases Φ^{lex} in the lex order. A strictly increasing sequence in a finite totally-ordered set is finite, so the algorithm terminates. Stability of the terminal partition follows as in the proof of Theorem 7.1. □

Theorem 12.6 (Lex-Slack Recovery)**.** *Under restrictions (A1)–(A3) and symmetric friend-oriented preferences with $\mathcal{F}(\varepsilon) \neq \emptyset$, every Φ^{lex}-maximiser P^{**} in $\mathcal{F}(\varepsilon)$ lies in the ε-constrained core under strict lex preferences.*

Proof. Φ^{lex} attains its maximum on the finite set $\mathcal{F}(\varepsilon)$ (totally ordered range). Suppose, towards a contradiction, that some coalition T blocks P^{**} via τ_{new} under strict lex preference: every $i \in T$ has $P' \succ_i P^{**}$. This implies (R4′). By Lemma 12.4, $\Phi^{\text{lex}}(P') >_{\text{lex}} \Phi^{\text{lex}}(P^{**})$, with $P' \in \mathcal{F}(\varepsilon)$ from (R3). This contradicts the lex-maximality of P^{**}. □

12.5 Why symmetry is essential

The argument in Lemma 12.4 relies on Lemma 5.1, which in turn relies on the symmetric doubling: each within-class pair contributes its weight twice to Φ. Under *asymmetric* friendship—where $j \in F(i)$ does not imply $i \in F(j)$—the friendship game becomes a directed-edge game without this doubling, and the strict-ascent identity does not hold. Concretely: a deviation can strictly improve every coalition member's friend-count while reducing the friend-count of a non-coalition agent whose admiration arc happened to point into the coalition. The Φ_F ascent argument then breaks.

This failure is not a technical artefact. The next section exhibits an explicit instance under asymmetric friendship in which the ε-constrained core is empty over an interval of ε, despite $\mathcal{F}(\varepsilon)$ being non-empty.

13 Emptiness under Asymmetric Friendship

We now construct an instance in which $\mathcal{F}(\varepsilon)$ is non-empty yet the ε-constrained core is empty over a sharp interval $\varepsilon \in [1/10, 2/5)$. The instance uses the original KKÖ domain—asymmetric friend-oriented preferences—and exhibits a clean three-regime phase structure as ε varies.

13.1 Why asymmetric friendship is the right setting

Under symmetric friendship, Theorems 11.2 and 12.6 guarantee non-emptiness of the ε-constrained core. The recovery argument relies on a single technical step: a strict-improvement blocking coalition implies a strict potential increase, contradicting maximality. Asymmetric friendship is the minimal escape route from this conclusion, because under asymmetry the potential-ascent inequality (1) no longer follows from (R4) alone: a non-coalition agent's admiration arc into the coalition can produce a within-class friendship loss that is not counted on the coalition side, breaking the symmetric doubling. The move-graph framework no longer applies directly, and we verify the emptiness phenomenon computationally.

13.2 The five-agent instance

Theorem 13.1 (Emptiness of the ε-constrained core under asymmetric friendship)**.** *There exists an instance with asymmetric friend-oriented preferences, fixed class sizes, and an interval*

$\varepsilon \in [1/10, 2/5)$ *for which* $\mathcal{F}(\varepsilon) \neq \emptyset$ *but the* ε*-constrained core is empty. The instance further exhibits a sharp three-regime structure:*

(i) $\varepsilon < 1/10$*:* $\mathcal{F}(\varepsilon) = \emptyset$*.*

(ii) $1/10 \leq \varepsilon < 2/5$*:* $\mathcal{F}(\varepsilon) \neq \emptyset$*,* ε*-constrained core empty.*

(iii) $\varepsilon \geq 2/5$*:* ε*-constrained core non-empty.*

Construction and verification. Instance. $N = \{r_1, r_2, b_1, b_2, b_3\}$, $R = \{r_1, r_2\}$, $B = \{b_1, b_2, b_3\}$, $p_R = 2/5$. Two classes of sizes $q_1 = 2$, $q_2 = 3$.

Asymmetric friendship. Each agent's friendship-out set:

$$F(r_1) = \{r_2\}, \quad F(r_2) = \{b_2\}, \quad F(b_1) = \{b_2\}, \quad F(b_2) = \{b_3\}, \quad F(b_3) = \{b_1\}.$$

Structural features: (i) $r_1 \to r_2$ is unreciprocated ($r_2 \not\to r_1$); (ii) the three blues form a directed friendship 3-cycle $b_1 \to b_2 \to b_3 \to b_1$; (iii) $r_2 \to b_2$ is unreciprocated cross-colour.

Fairness analysis. For a class of size q with r reds, the demographic deviation is $|r/q - 2/5|$. Possible $(r^{(1)}, r^{(2)})$ distributions of reds (with $r^{(1)} + r^{(2)} = 2$):

- $(0, 2)$: $|0 - 2/5| = 2/5$ (size 2); $|2/3 - 2/5| = 4/15$ (size 3); binding $2/5$.
- $(1, 1)$: $|1/2 - 2/5| = 1/10$ (size 2); $|1/3 - 2/5| = 1/15$ (size 3); binding $1/10$.
- $(2, 0)$: $|1 - 2/5| = 3/5$ (size 2); $|0 - 2/5| = 2/5$ (size 3); binding $3/5$.

Hence

- $\varepsilon < 1/10$: $\mathcal{F}(\varepsilon) = \emptyset$.
- $1/10 \leq \varepsilon < 2/5$: only $(1, 1)$ feasible; six partitions (choose 1 of 2 reds and 1 of 3 blues for the size-2 class).
- $2/5 \leq \varepsilon < 3/5$: $(1, 1)$ and $(0, 2)$ feasible; nine partitions.
- $\varepsilon \geq 3/5$: all three distributions feasible; ten partitions.

Claim 1: For $1/10 \leq \varepsilon < 2/5$**, every feasible partition is blocked.**

Every $(1, 1)$-partition has the form $(\{b_x, r_y\}, \{b_p, b_q, r_{3-y}\})$ for some $b_x \in \{b_1, b_2, b_3\}$, $r_y \in \{r_1, r_2\}$, and $\{b_p, b_q\} = \{b_1, b_2, b_3\} \setminus \{b_x\}$.

We exhibit a blocking coalition for each such partition. The directed friendship cycle $b_1 \to b_2 \to b_3 \to b_1$ implies: for any blue b_x in the size-2 class, there is a unique blue b_z in the size-3 class whose outward friend is b_x (i.e., $F(b_z) = \{b_x\}$). Consider the same-colour swap $b_x \leftrightarrow b_z$. Same-colour swaps preserve fairness by Proposition 3.2.

After the swap, b_x is in the size-3 class together with b_y (the unique blue with $F(b_x) = \{b_y\}$, namely $b_y = b_{x'}$ for $x' = x + 1 \mod 3$ identifying with the directed cycle). b_x's friend-count goes from 0 (her friend b_y was in the same class as her in the original partition only if she was in the size-3 class, but she was in the size-2 class) to either 1 (if b_y ends up in the size-3 class together with b_x, which happens by the structure of the swap) or higher. Strict gain in f_{b_x}.

For b_z: she moves from the size-3 class to the size-2 class. Her outward friend b_x has just moved out of the size-3 class, so f_{b_z} stays at zero. Her enemy count: in the size-3 class she had two non-friends; in the size-2 class she has one non-friend (r_y). So e_{b_z} strictly decreases. Under lex preferences with f-coordinate tied at zero, this is a strict improvement.

Hence the coalition $T = \{b_x, b_z\}$ blocks every $(1, 1)$-partition. Exhaustive computational enumeration confirms that all six $(1, 1)$-partitions admit such a blocking pair at $\varepsilon = 1/10$ (and hence at every $\varepsilon \in [1/10, 2/5)$, since these partitions are the only feasible ones).

Claim 2: For $\varepsilon \geq 2/5$, the partition $P^\dagger = \{\{b_1, b_3\}, \{b_2, r_1, r_2\}\}$ lies in the ε-constrained core.

Friend- and enemy-counts at $P^\dagger$:

agent	b_1	b_3	b_2	r_1	r_2
$f_i(P^\dagger)$	0	1	0	1	1
$e_i(P^\dagger)$	1	0	2	1	1

Justification: b_1's outward friend b_2 is in class 2, not class 1; b_3's outward friend b_1 is in class 1, same class; b_2's outward friend b_3 is in class 1, not class 2; r_1's outward friend r_2 is in class 2, same class; r_2's outward friend b_2 is in class 2, same class.

For any coalition T to strictly improve every member under strict lex preferences, each $i \in T$ must either (i) strictly gain a friend, or (ii) keep f_i constant and strictly reduce e_i.

b_3 *cannot strictly improve.* She already has $f = 1$ (her unique outward friend b_1 is in class) and $e = 0$ (minimum). No further friend is reachable, and e cannot decrease below zero. So $b_3 \notin T$.

r_1 *cannot strictly improve.* She has $f = 1$ (her unique friend r_2 in class) and $e = 1$. To strictly improve she must keep r_2 in class *and* reduce e. Her single enemy in class is b_2. To remove b_2 from class 2 without losing r_2, the reassignment must move b_2 out of class 2 while keeping r_1 and r_2 in class 2. Capacity preservation requires one student to move into class 2 from class 1 in exchange. The candidates are b_1 and b_3. Either way, b_2 is sent to a class without her outward friend b_3 (since b_3 would also need to move out for b_2 to end up next to b_3, but that would also lose r_2). So b_2's friend-count drops below its current value of zero, impossible. Alternatively, r_1 moves to class 1: she leaves r_2 and loses her friend, dropping f_{r_1}. Not strict improvement. Hence $r_1 \notin T$.

r_2 *cannot strictly improve.* Symmetric to r_1 (the structure is analogous: unique friend b_2 in class, single enemy r_1 in class). $r_2 \notin T$.

Hence $T \subseteq \{b_1, b_2\}$. With $|T| \geq 2$ for any capacity-preserving deviation, $T = \{b_1, b_2\}$. The only capacity-preserving reassignment is the swap $b_1 \leftrightarrow b_2$, producing $P' = \{\{b_2, b_3\}, \{b_1, r_1, r_2\}\}$. Check b_1 at P': in class $\{b_1, r_1, r_2\}$, with outward friend b_2 not in class, so $f_{b_1} = 0$ (unchanged); enemies in class are $\{r_1, r_2\}$, so $e_{b_1} = 2$, up from 1. b_1 strictly loses on the second lex coordinate; no strict improvement. Hence no blocking coalition exists, and $P^\dagger$ is in the ε-constrained core for all ε at which it is feasible, in particular $\varepsilon \geq 2/5$.

Phase boundaries. Combining feasibility analysis with Claims 1 and 2:

- $\varepsilon < 1/10$: regime (i), $\mathcal{F}(\varepsilon) = \emptyset$, infeasibility.
- $\varepsilon \in [1/10, 2/5)$: regime (ii), $\mathcal{F}(\varepsilon) \neq \emptyset$ (six $(1,1)$-partitions) but core empty (every partition blocked by Claim 1). *Homophilic squeeze.*
- $\varepsilon \geq 2/5$: regime (iii), $(0,2)$-partition $P^\dagger$ becomes feasible and lies in core (Claim 2). At $\varepsilon \geq 3/5$, $(2,0)$-partitions also enter the core.

□

Remark 13.2 (Cyclic blocking signature)**.** *The blocking deviations among the six* $(1,1)$*-partitions form a Condorcet-style cycle: each* $(1,1)$*-partition is dominated by another* $(1,1)$*-partition under strict lex preference, with the dominance arrows tracing the friendship cycle* $b_1 \to b_2 \to b_3 \to b_1$*. Cyclic blocking is the cooperative-game-theoretic signature of an empty core.*

Remark 13.3 (Price of fairness at the existence level)**.** *The welfare-based price of fairness studied in the companion paper measures how much* Φ *is sacrificed for demographic balance. Theorem 13.1 exhibits a qualitatively different cost: in the homophilic-squeeze regime,* ε*-fairness destroys core stability* as an existence question. *The welfare price of fairness is undefined when there is no core-stable partition to evaluate.*

Remark 13.4 (Why fixed sizes are essential)**.** *Without restriction (A2), the grand coalition N has demographic share exactly p_R and is always fair; it sits in the core by construction. The empty-core phenomenon requires that ε-fairness genuinely constrains the admissible partitions, which only happens when class capacities bind. This is also the realistic regime for school assignment.*

13.3 The full stability hierarchy

Corollary 13.5 (Full stability hierarchy)**.** *Combining the symmetric and asymmetric results:*

1. *Symmetric friendship (additive or lex), $\mathcal{F}(\varepsilon) \neq \emptyset$: ε-constrained core non-empty (Theorems 11.2, 12.6).*
2. *Asymmetric friendship: ε-constrained core may be empty, with a three-regime ε-phase structure on at least one instance (Theorem 13.1).*
3. *In both regimes under symmetric friendship: a fairly $s_{\max}$-coalitionally stable partition exists and is computed by Algorithm 2 in polynomial time (Theorems 7.1, 8.5).*

Fair $s_{\max}$-coalitional stability is the robust fallback: it is always achievable under symmetric friendship; the stronger ε-constrained core may fail to exist under asymmetric friendship.

14 Conclusion

The move-graph framework offers a unified treatment of coalitional refinement under ε-fairness. The central object G_T replaces permutation-cycle reasoning with the Eulerian decomposition of a directed multigraph; this design (a) eliminates case-splits in the cycle-suffices argument, (b) handles the pathological-cancellation phenomenon (Lemma 9.1) without algorithmic modification, and (c) admits clean extensions to friend-oriented preferences and to the KKÖ setting.

The embedding into KKÖ establishes a Welfare-Slack Recovery Theorem (Theorem 11.2) showing that the ε-constrained core under symmetric additive preferences is non-empty whenever a fair partition exists, and a parallel result under lex (Theorem 12.6). The within-group strategy-proofness result (Proposition 11.5) shows that the natural Φ-maximising mechanism is incentive-compatible against within-group withholding manipulations under a mutual-consent reporting convention.

The emptiness phenomenon and the corresponding three-regime ε-phase boundary (Theorem 13.1) require genuinely asymmetric friendship. This locates precisely where ε-fairness and KKÖ-style stability are incompatible: in the homophilic-squeeze regime of asymmetric friendship under tight ε. Outside this regime, the stability hierarchy is well-behaved: the strong ε-constrained core when ε is slack, fair $s_{\max}$-coalitional stability otherwise (Corollary 13.5).

Three directions remain open. First, extending the strict-ascent argument (Lemma 5.1) to asymmetric friendship would either establish move-graph-based existence results in that domain or rule them out structurally; the obstruction is the failure of the symmetric doubling underlying the proof. Second, a closed-form characterisation of the boundary between empty-core and non-empty-core regimes for asymmetric instances beyond the five-agent witness would yield structural insight into when KKÖ-stability survives fairness constraints. Third, the tractable computation of an ε-constrained SCC mechanism, analogous to fairlet-based fair clustering, would close the loop between this framework and the existing fair-clustering literature.

References

Aamand, A., Chen, J., Liu, A., Silwal, S., Sukprasert, P., and Vakilian, A. (2023). Constant approximation for individual preference stable clustering. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 43646–43661.

Ahmadi, S., Awasthi, P., Khuller, S., Kleindessner, M., Morgenstern, J., Sukprasert, P., and Vakilian, A. (2022). Individual preference stability for clustering. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 162:197–246.

Apt, K. R., de Keijzer, B., Rahn, M., Schäfer, G., and Simon, S. (2017). Coordination games on graphs. *International Journal of Game Theory*, 46(3):851–877.

Apt, K. R., Schäfer, G., and Simon, S. (2020). Coordination games on weighted directed graphs. *Mathematics of Operations Research*, 47(2):995–1025.

Aziz, H., Brandt, F., and Seedig, H. G. (2013). Computing desirable partitions in additively separable hedonic games. *Artificial Intelligence*, 195:316–334.

Bilò, V., Monaco, G., and Moscardelli, L. (2022). Hedonic games with fixed-size coalitions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 4834–4842.

Bogomolnaia, A. and Jackson, M. O. (2002). The stability of hedonic coalition structures. *Games and Economic Behavior*, 38(2):201–230.

Brandt, F., Bullinger, M., and Tappe, A. (2025). Single-deviation stability in additively separable hedonic games with constrained coalition sizes. arXiv:2510.12641.

Caragiannis, I., Kanellopoulos, P., and Kyropoulou, M. (2022). Quantifying core stability relaxations in hedonic games. arXiv:2412.01666.

Currarini, S., Jackson, M. O., and Pin, P. (2009). An economic model of friendship: Homophily, minorities, and segregation. *Econometrica*, 77(4):1003–1045.

Currarini, S., Jackson, M. O., and Pin, P. (2010). Identifying the roles of race-based choice and chance in high school friendship network formation. *Proceedings of the National Academy of Sciences*, 107(11):4857–4861.

Dimitrov, D., Borm, P., Hendrickx, R., and Sung, S. C. (2006). Simple priorities and core stability in hedonic games. *Social Choice and Welfare*, 26(2):421–433.

Echenique, F. and Fryer Jr, R. G. (2007). A measure of segregation based on social interactions. *Quarterly Journal of Economics*, 122(2):441–485.

Fioravanti, S., Flammini, M., Kodric, B., and Varricchio, G. (2023). ε-fractional core stability in hedonic games. In *Proceedings of the 37th Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Fioravantes, F., Monnot, J., and Pickl, S. (2025). Parameterized welfare maximization in additively separable hedonic games with bounded coalition sizes. Working paper.

Klaus, B., Klijn, F., and Özbilen, S. (2025). Core stability and strategy-proofness in hedonic coalition formation problems with friend-oriented preferences. *Games and Economic Behavior*, 154:16–52.

Bullinger, M. and Kraiczy, S. (2025). Deviation dynamics in cardinal hedonic games. arXiv:2511.11531.

Kuniavsky, S. and Smorodinsky, R. (2014). Equilibrium and potential in coalitional congestion games. arXiv:1111.3933.

Levinger, A., Hazon, N., and Azaria, A. (2024). Welfare maximization and group stability in additively separable hedonic games with coalition-size bounds. Working paper.

Monderer, D. and Shapley, L. S. (1996). Potential games. *Games and Economic Behavior*, 14(1):124–143.

Wright, M. and Vorobeychik, Y. (2016). Mechanism design for team formation. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 1050–1056.