UNIVERSITY OF NAPLES FEDERICO II

DOCTORAL THESIS

Three Essays on Matching theory

Author
Andrei Belykh

Supervisor
Prof. Michele Lombardi

Supervisor
Prof. Pietro Salmaso

PhD Director
Prof. Marco PAGANO

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy in Economics - XXXVII Cycle at the

Naples School of Economics Department of Economics and Statistical Sciences

Abstract

This dissertation investigates how collective allocation and coalition-formation mechanisms perform when individual preferences exhibit interdependencies, and when classical notions of stability become unattainable or insufficient. The first chapter studies school choice in environments with peer-dependent preferences, where each student's utility depends not only on the assigned school but also on the composition of peers. Such externalities eliminate the possibility of stable assignments. Nevertheless, the Rank-Minimizing mechanism retains its Non-Obvious Manipulability property in this richer environment, which is the main result of the first chapter.

The second chapter examines the weak strategic properties of school choice mechanisms more broadly. It shows that, for unanimous mechanisms, regret-free truth-telling implies non-obvious manipulability, but not vice versa. The RM-SD mechanism serves as a concrete example in which non-obvious manipulability holds without guaranteeing regret-free truthful reporting. Taken together, these results establish a more complete hierarchy of strategic properties, ranging from obvious strategy-proofness to mechanisms that are clearly manipulable.

The third chapter turns to decentralized insurance markets and analyzes the endogenous formation of reinsurance pools as a hedonic coalition game. Insurers differ in capital endowments and probabilities of bankruptcy, and a solvency regulation parameter limits the aggregate risk exposure of any coalition. Under mild homogeneity of risk characteristics and strictly monotone preferences for more reliable partners, the induced game admits a unique core-stable partition: the stable coalitions consist of consecutive sets of the most reliable firms. This result characterizes how capital regulation shapes coalition structure and identifies the endogenous organization of risk-sharing arrangements.

Taken together, the dissertation advances our understanding of matching and coalition-formation mechanisms in environments with externalities and strategic complexity. It highlights settings in which stable assignments or core outcomes fail to exist and classical incentive compatibility notions weaken, yet mechanisms can still deliver behaviorally robust and economically meaningful allocations.

Contents

Chapter 1

Non-Obvious Manipulability of the Rank-Minimizing Mechanism in School Choice Problem with Peers

Andrei Belykh

This paper studies the performance of the Rank-Minimizing (RM) mechanism in a school choice environment with peer-dependent preferences. In this setting, each student's satisfaction depends not only on the assigned school but also on the peers allocated to it, which introduces externalities and makes stability unattainable. The analysis shows that the RM mechanism remains non-obviously manipulable: while agents may have theoretical incentives to misreport, such manipulations are not cognitively transparent. The result demonstrates that the mechanism preserves its behavioral robustness and efficiency in environments with peer effects, providing a foundation for further exploration of efficiency-based designs when stability is infeasible.

Key Words school choice, peer effects, rank-minimizing mechanism, non-obvious manipulability, strategy-proofness, mechanism design, matching with externalities

1 Introduction

The school choice problem and the design of appropriate allocation mechanisms remain central issues in both economic theory and policy applications. As noted by Troyan (2023), a natural measure of the performance of a school choice mechanism is the average rank of the assigned schools from the students' reported preferences. This measure directly reflects how well the mechanism performs in terms of welfare, as perceived by the participants. Ortega and Klein (2023) further demonstrate that the rank-minimizing algorithm achieves the highest efficiency with respect to this metric, outperforming standard mechanisms such as the Deferred Acceptance (DA), Top Trading Cycles (TTC), and the Efficiency-Adjusted Deferred Acceptance Mechanism (EADAM).

Building on this perspective, the classic result of Abdulkadiroglu and Sonmez (2003) establishes that no school choice mechanism can be simultaneously Pareto efficient and free of justified envy, the latter corresponding to the property of stability. This fundamental incompatibility has motivated subsequent research seeking to reconcile efficiency and fairness. Kesten (2010) proposed the Efficiency-Adjusted Deferred Acceptance Mechanism (EADAM), which allows students to voluntarily waive certain priorities without reducing their own welfare, thereby potentially improving aggregate efficiency. Experimental evidence by Cerrone (2024) supports the practical relevance of EADAM, showing that it achieves higher efficiency and induces greater truth-telling rates compared to the classical Deferred Acceptance (DA) mechanism. However, Cerrone (2024) also notes that this improvement in efficiency comes at the cost of increased exposure to strategic manipulation. Further contributions by Abdulkadiroglu (2003, 2008) emphasize both the limitations of existing assignment plans and potential refinements of the DA mechanism, including procedures that allow students to influence outcomes in tie-breaking situations.

Subsequent research has continued to explore the trade-off between stability and efficiency by proposing various refinements of existing school choice mechanisms. Alcalde (2011) introduces a modification of the Boston mechanism aimed at improving both Pareto efficiency and stability under specific preference domains. Haeringer (2009, 2011) investigates the Student-Optimal Stable mechanism and the Top Trading Cycles (TTC) mechanism, identifying necessary and sufficient conditions under which these mechanisms achieve stable and efficient outcomes. Chen (2006) provides experimental evidence comparing the Boston, Deferred Acceptance (DA), and TTC mechanisms, showing that DA and TTC outperform the Boston mechanism in terms of efficiency, particularly when participants are informed about the mechanisms' incentive properties. More recently, Ortega and Klein (2023) compare the outcomes generated by the prominent school choice mechanisms EADAM, DA, and TTC with those produced by the Rank-Minimizing (RM) mechanism. Their findings indicate that the RM mechanism yields systematically higher allocative performance, enhances the welfare of the least-advantaged students, and reduces the incidence of justified envy relative to the TTC mechanism. As an alternative to choosing a mechanism, a solution can be considered where not only the mechanism changes, but also the whole problem. If students are allowed to choose not only the school, but also with whom they would like to study at this school, then the problem ceases to have a stable solution. The analysis of cases with a lack of stability in the presence of externalities is presented in the Bando, Ryo and Shigeo (2016). It also examines in detail the lack of stability for the case of the School Choice problem with colleagues.

The introduction of externalities into the school choice problem was first formalized by Dutta and Masso (1997), who proposed an extended framework commonly referred to as the *School Choice Problem with Colleagues* (peers). In their model, each student's preferences depend not only on the school itself but also on the peers with whom they are assigned. Dutta and Masso (1997) show that a stable matching exists only under restrictive preference domains, specifically when all students exhibit college-lexicographic preferences. When preferences are unrestricted, stability generally fails to exist. Revilla (2007) reinforces this finding by demonstrating that, in the presence of arbitrary preference structures, the core may be empty. A related but distinct line of research is developed by Dur and Wiseman (2019), who incorporate preferences over a single peer into the standard school choice model. Even in this simplified environment, stable outcomes are not guaranteed. While the Dur and Wiseman (2019) setting can be viewed as an extension of the classical framework, the Dutta and Masso (1997) model fundamentally alters its structure by explicitly embedding externalities into agents' preferences. In environments where stability cannot be assured, it becomes natural to shift the focus toward efficiency, with the Rank-Minimizing (RM) mechanism emerging as a particularly suitable candidate for such settings.

Incorporating preferences over both schools and peers is not only conceptually relevant but also empirically well motivated. A substantial body of evidence documents that the composition of one's peer group has a meaningful impact on academic outcomes. Hanushek, Kain, Markman, and Rivkin (2003) show that exposure to higher-ability classmates raises individual achievement in primary and secondary schools. Duflo, Dupas, and Kremer (2011), using a randomized tracking intervention, demonstrate that grouping students by ability can improve learning, partly through peer interactions. Mehta (2018) provides further evidence at the college level, showing that students' academic effort responds to the study habits of their peers. These findings indicate that peer effects are both sizable and behaviorally salient, suggesting that models of school choice that allow students to value the identity and characteristics of their potential classmates have clear empirical grounding.

While the question of stability becomes intractable in environments with peer-related externalities, the problem of strategic behavior by participants remains central. Once efficiency is chosen as the guiding criterion, it becomes crucial to understand whether the underlying mechanism preserves desirable incentive properties. The literature on manipulability in school choice mechanisms (Kesten, 2012; Mennle, 2016; Troyan, 2023) addresses precisely this issue, showing that even mechanisms designed for efficiency, such as the Rank-Minimizing (RM) algorithm, can be susceptible to manipulation. However, as emphasized by Troyan and Morrill (2020), the RM mechanism is non-obviously manipulable, meaning that while manipulation is theoretically possible, it is not cognitively apparent to boundedly rational agents.

While the absence of stability in environments with peer-related externalities challenges traditional approaches to the school choice problem, another important dimension concerns the incentive properties of allocation mechanisms. Once stability is no longer attainable, the question naturally shifts toward whether efficiency-oriented mechanisms can retain desirable behavioral characteristics. In particular, it becomes relevant to ask whether the Rank-Minimizing (RM) mechanism preserves its positive properties in such an extended environment, most notably the property of *non-obvious manipulability*.

The issue of manipulability in school choice mechanisms has been extensively analyzed in the literature. Kesten (2012) investigates manipulation through the use of capacities and pre-arranged matches, while Mennle (2016) studies first-choice maximality and the strategic vulnerabilities that accompany it. Building on these insights, Troyan (2023) examines the Rank-Minimizing mechanism within the classical school choice framework and finds that, although the mechanism is not strategy-proof, it remains *non-obviously manipulable*. This means that while agents may have incentives to misreport preferences, identifying profitable manipulations requires cognitive sophistication that exceeds bounded rationality.

The conceptual foundation for this idea was established by Troyan and Morrill (2020), who introduced *non-obvious manipulability* as a behavioral relaxation of strategy-proofness. Mechanisms with this property are formally manipulable but difficult to exploit because the benefits of deviation are neither immediate nor transparent to agents. This behavioral robustness makes such mechanisms particularly appealing in practical environments where full rationality cannot be assumed.

In this paper, we extend the classical school choice framework by allowing students to express preferences not only over schools but also over the peers with whom they wish to study. The presence of such peer-related preferences introduces externalities, as each student's utility from attending a particular school depends on the enrollment decisions of other students. Consistent with the existing literature, schools are assumed to have limited capacities, and students hold strict preferences over both schools and possible peer groupings. Within this environment, we analyze the performance of the Rank-Minimizing (RM) mechanism, which assigns students to schools so as to minimize the average rank of the resulting allocation. The main result of the paper establishes that the RM mechanism retains the property of *non-obvious manipulability* even in this extended setting with externalities.

The analysis is conducted within the formal framework developed by Dutta and Masso (1997). While this framework is not directly implementable in real-world school choice systems—since it assumes that each student can rank all possible combinations of schools and peer groups—it provides a valuable theoretical foundation. Studying mechanisms in this broad setting allows us to identify structural properties that may persist in more realistic, restricted environments. In particular, the model can be meaningfully narrowed to cases where a student's preferences depend only on a small subset of peers (for instance, two or three classmates), while preserving the essential features of the mechanisms under consideration. This approach appears more tractable than extending the model of Dur and Wiseman (2019), which, although simpler in its current form, would require significant technical modifications to accommodate richer peer interactions. Thus, although the present analysis remains theoretical, it offers useful insights and policy-relevant implications for designing efficient school choice mechanisms when stability cannot be guaranteed.

It is important to emphasize that the result obtained in this paper differs fundamentally from that of Troyan (2023), as the model considered here cannot be reduced to the classical school choice framework he analyzes. While Kominers (2010) demonstrates that such a reduction is theoretically possible, it requires highly restrictive and unrealistic assumptions. Absent these restrictions, the two models coincide only in degenerate cases, such as when each school has a single seat or when there is only one student, which are of limited theoretical or practical relevance. Consequently, the present study extends the analysis of the Rank-Minimizing mechanism to a qualitatively richer environment characterized by peer-dependent preferences and externalities.

To maintain analytical tractability and comparability with the existing literature, we impose a lower bound on the number of students in the model. This assumption excludes trivial cases involving only one or two students, which would be of limited theoretical interest. It also facilitates meaningful comparisons with related algorithms developed for similar environments, such as the one proposed by Kondratiev and Nesterov (2021). Furthermore, throughout the analysis, we abstract from school priorities, both within the mechanism and in the computation of the average rank. This simplification is standard in the school choice literature and allows us to focus more clearly on the interaction between efficiency and manipulability in environments with externalities.

The next section introduces the formal model and establishes the notation and definitions used throughout the paper.

2 Model

Let C be a set of schools and S be a set of students with $|S| = N > 2$ and $|C| = K$. Each school $c \in C$ has a strict priority ordering $\succ_c$ over 2^S. We denote the set of all subsets of students that contain s as: $S_s \equiv \{S' \in 2^S | s \in S'\}$.

Each student s has a strict preference ordering $\succ_s$ over possible options $O_s \equiv (C \times S_s) \cup \{(\emptyset, \emptyset)\}$. Any feasible ordering of the options is possible. We denote the option $(\emptyset, \emptyset)$ to be unmatched (the outside option), but it is not allowed to have this option as the best choice. Each school has a capacity q_m which denotes the maximum number of students that can be assigned to this school. We use the following assumption: $\sum_{m=1}^{K} q_m = N$, with the economic intuition behind this assumption being the following: on the one hand, there are enough places for all students, and on the other hand, we do not want to have wasted resources (ex-ante for sure unoccupied places).

In our analysis, we impose several mild but conceptually important restrictions on the environment to ensure that the model captures genuine peer-related externalities and remains analytically coherent.

First, we assume that each school c has a capacity of at least two $\forall m \in R, q_m \geq 2$. This restriction follows naturally from the Dutta and Masso (1997) framework, where preferences are defined over combinations of schools and peer groups. When a school has only one seat, no peer interaction can occur, and the environment collapses to the classical school choice model without externalities. Excluding single-seat schools avoids such degenerate cases and keeps the analysis focused on situations in which peer effects are operative.

Second, we restrict the feasible set of allocations by excluding options in which a student would be the only enrolled student at a school. These configurations are both unrealistic and inconsistent with the motivation for introducing peer-dependent preferences.

Third, we assume that the outside option, representing for instance non-assignment or opting out of the system, cannot appear as the top-ranked alternative in a student's preference ordering. This assumption prevents trivial preference profiles in which participation in the allocation mechanism is irrelevant or dominated by self-exclusion. It also ensures that the model remains informative about the relative performance of assignment mechanisms under active participation.

Taken together, these restrictions maintain the analytical focus on nontrivial environments in which externalities among students are meaningful, while excluding degenerate or behaviorally irrelevant configurations. They also align the setting with institutional realities, where schools typically accommodate multiple students and participation in the allocation process is compulsory or at least desirable.

Let P_s be a complete preference domain for student s, and let $P_S = P_{s_1} \times P_{s_2} \times \ldots \times P_{s_N}$. At the same time $\succ_S = (\succ_{s_1}, \succ_{s_2}, \ldots, \succ_{s_N}) \in P_S$. From now we refer to the School Choice Problem with Peers (or economy) as tuple $\langle S, C, P_S \rangle$. The primitives introduced allow us to use the following definition:

Definition 1: A (deterministic) allocation $\boldsymbol{\alpha}: C \cup S \to 2^S \cup \{(2^S \times C)\} \cup \{(\emptyset, \emptyset)\}$ is a mapping that satisfies:

1) $\alpha(c) \in 2^S \; \forall c \in C$
2) $\alpha(s) \in (C \times S_s) \cup \{(\emptyset, \emptyset)\} \; \forall s \in S$
3) $s \in \alpha(c) \to \alpha(s) = (c, \alpha(c)) \; \forall s \in S \; \forall c \in C$
4) $\alpha(s) = (c, S') \in C \times S_s \to \alpha(c) = S' \; \forall s \in S' \subseteq S$

Definition 2: The rank of allocation α according to s_i's preferences is $r_{s_i}(\alpha(s_i)) = |o' \in O_s : o' \succsim_s \alpha(s_i)|$.

This definition introduces the ordinal welfare measure at the individual level. It establishes how each student evaluates the assigned alternative, providing the basic metric used to aggregate welfare and to compare allocations across mechanisms.

We consider **A** as the set of all possible allocations.

Definition 3: A random allocation $\mu: \boldsymbol{A} \to [0,1]$ is a probability distribution over **A**, where $\sum_{\alpha \in A} \mu(\alpha) = 1$

This extends the space of outcomes to probability distributions over deterministic matchings. It allows the mechanism to incorporate randomization and is essential for analyzing manipulability when outcomes are lotteries.

Let $\boldsymbol{M}$ be the set of random allocations.

The average rank of any matching then is defined as follows:

Definition 4: The average rank of any deterministic allocation $\boldsymbol{\alpha}$ is $\overline{r(\alpha)} = \frac{1}{N} * \sum_S r_S(\alpha)$

We extend preferences of students $\succ_s$ over the $\boldsymbol{A}$ using common notation: $\forall \alpha \in A, s \in S\ (\alpha' \succsim_s \alpha \leftrightarrow r_s(\alpha'(s)) \leq r_s(\alpha(s)))$. This means that to compare two allocations for any agent it is enough to compare the ranks of options that the agent gets from these allocations.

To formalize the efficiency benchmark, we proceed as follows.

The following definitions establish the conceptual and stochastic framework used in the remainder of the analysis. They identify the set of deterministic assignments that minimize the average rank and extend it to the space of random allocations whose support consists exclusively of such efficient outcomes. Introducing the notion of full-support random assignments provides a neutral and technically complete tie-breaking rule, ensuring that randomization preserves both efficiency and comparability across mechanisms. Together, these elements form the analytical foundation for the subsequent study of the Rank-Minimizing mechanism and its incentive properties.

Definition 5: Let $\bar{A}(\succ_S) = \{\alpha \in A: \bar{r}(\alpha) \leq \bar{r}(\alpha')\ \forall \alpha' \in A\}$ is the *set of rank-minimizing deterministic allocations*.

The set $\bar{A}(\succ_S)$ is not empty for all $\succ_S$ because A is finite. $\bar{A}(\succ_S)$ can have more than one element.

It is necessary to incorporate random allocations in order to handle tie-breaking situations that arise when multiple deterministic allocations yield the same minimal average rank.

Definition 6: A random allocation μ is a *rank-minimizing random allocation* with respect to $\succ_S$ if $\alpha \notin \bar{A}(\succ_S) \rightarrow \mu(\alpha) = 0$.

In words, μ places positive probability weight on only those deterministic allocations that are rank-minimizing given the preference profile $\succ_S$. We denote $\bar{\boldsymbol{M}}(\succ_S)$ the set of rank-minimizing random allocations.

Definition 7: When $\mu(\alpha) > 0$ for all $\alpha \in \bar{A}(\succ_S)$ we call μ a *full-support rank-minimizing random allocation*.

To illustrate the structure of each agent's preferences, consider an example. We have $S = \{s_1, s_2, s_3\}, C = \{c_1, c_2\}$ and a preference profile of student s_1 below:

$\succ_{s_1}$	Rank
$(c_1, \{s_1, s_2, s_3\})$	1
$(c_2, \{s_1, s_2, s_3\})$	2
$(c_1, \{s_1, s_2\})$	3
$(c_2, \{s_1, s_2\})$	4
$(\emptyset, \emptyset)$	5
…	…

For student s_1, the best option is to be at school c_1 with student s_2 and s_3. This option has a rank equal to one. At the same time, if student s_1 is at school c_1 without his friend s_3, then this is only the third most preferred option, which has a rank of 3. So if we have a matching $\alpha_1 = (\{c_1, s_1, s_2\}, \{(\emptyset, \emptyset), s_3\})$ and $\alpha_2 = (\{c_1, s_1, s_2, s_3\})$, then the second one is more preferable for the student s_1. Additionally, it is important to highlight that the length of the preference list is the same for every student.

Definition 8: A stochastic mechanism (an algorithm) is a function $\varphi\colon P^S \to M$.

For each preference profile $\succ_S \in P^S$ mechanism φ returns the random allocation $\mu \in M$. For any deterministic allocation $\alpha \in A$, we wright $\varphi(\succ_S)\ (\alpha)$ to denote the probability that the random allocation $\mu = \varphi(\succ_S)$ places on α.

We consider the rank-minimizing mechanism, which we define as follows.

Definition 9: φ^{RM} is the rank-minimizing (RM) mechanism if $\varphi^{RM}(\succ_S) \in \bar{\boldsymbol{M}}\ \forall \succ_S \in P^S$.

To complete the efficiency framework and prepare the incentive analysis, we fix the tie-breaking rule and the notation for random outcomes. The full-support rank-minimizing mechanism selects a distribution that assigns positive probability to every deterministic minimizer of the average rank. This neutral tie-breaking removes arbitrary selection within the efficiency set and ensures that all efficient assignments are taken into account in the analysis. We then formalize the support of a random distribution in order to speak precisely about outcomes that occur with positive probability. Finally, we define the best and worst attainable ranks for a student relative to a given distribution. These extremal ranks serve as the basis for the subsequent best-case and worst-case comparisons that underpin the analysis of non-obvious manipulability.

Definition 10: If $\varphi^{RM}(\succ_S) = \mu$ is a full-support rank-minimizing allocation $\forall \succ_S \in P^S$ then φ^{RM} is the full-support rank-minimizing (RM) mechanism.

A natural way to break ties is to select uniformly at random from the entire set $\bar{A}(\succ_S)$ of deterministic allocations that minimize the average rank. In this case we have full-support uniform rank-minimizing (URM) mechanism.

Definition 11: Given a random allocation $\mu \in M$, let $\text{supp}(\mu) = \{\alpha \in A\colon \mu(\alpha) > 0\}$ be the support of μ.

Definition 12: We denote the worst possible rank as $\overline{\rho_s}(\mu) = \max\limits_{\alpha \in \text{supp}(\mu)} r_s\big(\alpha(s)\big)$ and the best possible rank as $\underline{\rho_s}(\mu) = \min\limits_{\alpha \in \text{supp}(\mu)} r_s\big(\alpha(s)\big)$.

In the next part of the analysis, we turn to the incentive properties of the mechanisms, examining whether they are strategy-proof or susceptible to manipulation. For mechanisms that admit manipulation, we further distinguish between obvious and non-obvious forms, following the behavioral framework introduced by Troyan and Morrill (2020). This distinction is developed through a best- and worst-case comparison of attainable outcomes under truthful and non-truthful reports, allowing us to evaluate the cognitive accessibility of profitable deviations within the Rank-Minimizing mechanism.

3 Main Result

As discussed in the introduction, one of the desired properties of the mechanism is that the mechanism is strategy-proof.

To compare random assignments in an ordinal environment, we adopt stochastic dominance as the benchmark ordering over distributions of allocations.

Definition 13: (Stochastic Dominance)

Let $\succ_s \in P_S$ be the preferences of agent s, and let p and q be two *random allocation.*

Then p is said stochastically dominate q with respect to $\succ_s$, written $p \succsim_s^{SD} q$ if and only if for every $\alpha \in A$:

$$\sum_{\alpha' \succsim_s \alpha} p(\alpha') \geq \sum_{\alpha' \succsim_s \alpha} q(\alpha')$$

where $\alpha' \succsim_s \alpha$ means that agent s weakly prefers allocation α' to α under his preference ordering $\succ_s$.

The intuition is the following:

- $p \succsim_s^{SD} q$ means that, for every cutoff object, the probability of receiving an object ranked at least as good as that cutoff is no smaller under p than under q.
- If the inequality is strict for at least one allocation α, we write $p \succ_s^{SD} q$, meaning that p strictly stochastically dominates q.

Using this ordering, we define strategy-proofness as the requirement that truthful reporting weakly dominates any alternative report for every student.

Definition 14: A stochastic mechanism φ is said to be strategy-proof if, for every agent i, every pair of preference profiles $\succ_i, \succ_{-i}$, and any possible report $\succ_i'$, the following holds:

$$\varphi(\succ_i, \succ_{-i}) \succsim_s^{SD} \varphi(\succ_i', \succ_{-i})$$

A matching algorithm is said to be strategy-proof if, for every specification of agents' preferences, truthful reporting constitutes a dominant strategy for each participant. The mechanism considered here does not necessarily satisfy this property, since agents may have incentives to misreport their preferences. We therefore formalize manipulation as a profitable deviation, that is, a misreport that yields a distribution strictly preferred by the student according to stochastic dominance.

Definition 15: For any agent i with preferences $\succ_i \in P_S$, $\succ_i' \neq \succ_i \in P_S$ is a manipulation of mechanism φ if $\exists \succ_{-i} \in P_S$ $\varphi(\succ_i', \succ_{-i}) \succ_s^{SD} \varphi(\succ_i, \succ_{-i})$

If some agent i has a manipulation, then φ is manipulable.

Troyan (2022) observes that obvious manipulations correspond to those that can be identified by agents with bounded cognitive abilities, as characterized by Li (2017). We now state the following proposition.

Proposition 1: Full-support rank-minimizing (RM) mechanism is not strategy proof for the case of the school choice problem with peers.

Proof:

To establish the result, it suffices to provide an example in which a profitable manipulation exists. We consider the full-support Rank-Minimizing (RM) mechanism and illustrate the argument using the economy $\langle S, C, P_S \rangle$. $S = \{s_1, s_2, s_3\}, C = \{c_1\}$. The school capacity is $q_{c_1} = 3$.

The true preferences P_S are presented below:

$\succ_{s_1}$	$\succ_{s_2}$	$\succ_{s_3}$
$(c_1, \{s_1, s_2\})$	$(c_1, \{s_2, s_3\})$	$(c_1, \{s_3, s_1\})$
$(\emptyset, \emptyset)$	$(\emptyset, \emptyset)$	$(\emptyset, \emptyset)$
$(c_1, \{s_1, s_3\})$	$(c_1, \{s_2, s_1\})$	$(c_1, \{s_3, s_2\})$
$(c_1, \{s_1, s_2, s_3\})$	$(c_1, \{s_1, s_2, s_3\})$	$(c_1, \{s_1, s_2, s_3\})$

Under the full-support RM mechanism, there exist three deterministic allocations with an average rank of 2, which is the minimal attainable average rank among all feasible allocations.

$\alpha_1 = (\{c_1, s_1, s_2\}, \{s_3\}), \alpha_2 = (\{c_1, s_2, s_3\}, \{s_1\}\,), \alpha_3 = (\{c_1, s_3, s_1\}, \{s_2\}\,)$

For student s_1 we have: $\mu(\alpha_1), \mu(\alpha_2), \mu(\alpha_3) > 0$.

Let us say s_1 wants to improve his position. We can consider a situation in which s_1 reports preferences:

$\succ'_{s_1}$
$(c_1, \{s_1, s_2\})$
$(c_1, \{s_1, s_2, s_3\})$
$(\emptyset, \emptyset)$
$(c_1, \{s_1, s_3\})$

Now according to RM we get α_1 with the average rank 2 and all other matchings have higher average rank. For student s_1 we have:

$$\sum_{\alpha_1 \succsim_{s_1} \alpha_1} \mu'(\alpha_1) > \sum_{\alpha_1 \succsim_{s_1} \alpha_1} \mu(\alpha_1) \text{ because } 1 > 1 - \mu(\alpha_2) - \mu(\alpha_3) \text{ so}$$
$$\varphi(\succ'_{s_1}, \succ_{-s}) \succ^{SD}_{s_1} \varphi(\succ_{s_1}, \succ_{-s}). \; \square$$

The fact that the mechanism can be manipulated does not mean that it can be done obviously.

Definition 16: A mechanism φ is not obviously manipulable if, for any agent s with preferences $\succ_s$ and any $\succ'_s \neq \succ_s$ the following are true:

1) $\max\limits_{\succ_{-s}} \overline{\rho_S}(\,\varphi(\succ_s, \succ_{-s})) \leq \max\limits_{\succ_{-s}} \overline{\rho_S}(\,\varphi(\succ'_s, \succ_{-s}))$
2) $\min\limits_{\succ_{-s}} \underline{\rho_S}(\,\varphi(\succ_s, \succ_{-s})) \leq \min\limits_{\succ_{-s}} \underline{\rho_S}(\,\varphi(\succ'_s, \succ_{-s}))$

If 1) or 2) is not true than φ is obviously manipulable.

The interpretation of these two conditions captures the essence of non-obvious manipulability. Together, they ensure that there is no universal profitable deviation whereby an agent can improve their outcome through misreporting. The first condition guarantees that, across all possible preference profiles of other agents, the worst attainable rank under a misreport is never better than the worst attainable rank under truthful reporting. Similarly, the second condition states that the best attainable rank under a misreport is never better than the best attainable rank under truthful reporting.

Before turning to the main result, we introduce a lemma that provides a key intermediate step. Although technical in nature, it conveys an important economic intuition about the structure of incentives under the Rank-Minimizing mechanism.

Lemma 1: For any $s \in S$ with $\succ_s \in P_S$ and $o \in O_s$ there exists $\succ^*_{-s} \in P_{-s}$ such that $\alpha \in \bar{A}(\succ_s, \succ^*_{-s})$ with $\alpha(s) = o$.

The intuition behind Lemma 1 is that, for any given student and any feasible option, it is possible to construct the preferences of the remaining students so that the selected student obtains that option with a positive probability. The formal proof of Lemma 1 is provided in Appendix A. Building on this result, we now proceed to the main theorem.

Theorem 1: RM algorithm (mechanism) with full support is not obviously manipulable for the case of the school choice problem with peers.

The proof of Theorem 1 is in the Appendix B.

The main result demonstrates that the full-support Rank-Minimizing mechanism remains non-obviously manipulable in the school choice problem with peer-dependent preferences. This finding extends the existing literature by showing that the behavioral robustness of the mechanism persists even in environments characterized by externalities and the absence of stability. In what follows, we discuss the implications of this result for the design of allocation mechanisms and the potential directions for future research.

4 Discussions and Conclusion

The present paper extends the classical school choice framework by introducing peer-dependent preferences and analyzing the performance of the Rank-Minimizing (RM) mechanism in such an environment. The main theoretical contribution lies in showing that the RM mechanism remains non-obviously manipulable even when stability is unattainable due to externalities. This result highlights the behavioral robustness of the mechanism and suggests that efficiency-oriented designs can preserve desirable incentive properties beyond the classical setting.

From a broader perspective, the analysis contributes to the literature on mechanism design under bounded rationality. The finding that the RM mechanism is non-obviously manipulable indicates that, although formal incentives for misreporting exist, they are not easily exploitable by agents with limited cognitive capacity. This insight strengthens the case for considering the RM mechanism as a viable policy instrument when the planner values efficiency more than stability, as in large assignment markets or education systems emphasizing welfare maximization.

The framework of Dutta and Massó (1997) used here is intentionally general and not directly implementable in real-world school assignment systems. Nevertheless, its analytical richness allows us to identify structural features of mechanisms that may persist under more realistic restrictions. Future research should focus on narrowing this model to settings where students' preferences depend only on a limited number of peers. Such refinements would enable both theoretical extensions and empirical applications, bridging the gap between abstract mechanism design and practical implementation.

Overall, the findings of this paper demonstrate that the RM mechanism maintains its behavioral and efficiency properties even in complex environments with externalities.

This result encourages further exploration of efficiency-based allocation mechanisms in domains where stability is either unattainable or normatively undesirable. Future work could investigate hybrid mechanisms that combine rank minimization with partial stability or priority structures, as well as experimental evaluations of how agents behave under RM in realistic choice environments.

Appendix A

Lemma 1.

For any $s \in S$ with $\succ_s \in P_s$ and $o \in O_s$ there exists $\succ_{-s}^* \in P_{-s}$ such that $\alpha \in \bar{A}(\succ_s, \succ_{-s}^*)$ with $\alpha(s) = o$.

The proof plan will include the following steps:

I) Classification of all possible options available to each student

II) Proof

Proof of the Lemma 1:

I.

For any s_i, we can summarize all possible options by presenting them as three possible alternatives:

a)

$\boxed{(c_i, \{s_i, S_{s_i-i}\})}$

This option means that s_i is at c_i school with peers, and there are no additional available places at this school. The set S_{s_i-i} consists of students who are colleagues with s_i at a given school.

$S_{s_i} = S_{s_i-i} \cup \{s_i\}$

b)

$\boxed{(c_i, \{s_i, S_{s_i-i}, \emptyset\})}$

This option means that s_i is at c_i school with peers, and there are available places for new students in this school: one or more.

c)

$\boxed{(\emptyset, \emptyset)}$

Finally, s_i has the option not to be in any public school.

It is important to recall that, in our analysis, we exclude the possibility of a student being the sole enrollee at a school, as such configurations are unrealistic and inconsistent with the notion of peer-dependent preferences. Similarly, we rule out preference profiles in which the outside option appears as the top-ranked alternative. This restriction eliminates degenerate cases in which participation in the assignment mechanism becomes irrelevant and ensures that the analysis focuses on environments where students actively compete for school placements within socially meaningful peer groups.

II.

To proof the Lemma we need to show the existence of $\alpha \in \bar{A}(\succ_s, \succ_{-s}^*)$ which is true if for all $\alpha' \neq \alpha$ we have $\overline{r(\alpha')} \geq \overline{r(\alpha)}$.

a) <u>Suppose we want to give student s_i option</u> $o = (c_i, \{s_i, S_{s_i-i}\})$.

Let us consider the preferences $\succ_{-s}^*$ of the following type: there are $q_{c_i} - 1$ students for whom the option when they are all together with s_i in school c_i has a rank of 1. Let us denote the set of these students S_{s_i-i} and $S_{s_i} = S_{s_i-i} \cup \{s\}$. For the rest, $S' = S/S_{s_i}$ we have preference profiles such that they occupy all available places in the remaining $K - 1$

schools and this placement has the first rank for each of them. Below is a schematic description of the preferences of students from these two sets.

$\succ_s, s \in S_{s_i-i},$
$(c_i, \{s_i, S_{s_i-i}\})$
...
$to\ be\ in\ c_i\ without\ s_i$
$to\ be\ in\ c_i\ with\ s_i$
$(c_i, \{S_{s_i-i}\})$
$(\emptyset, \emptyset)$

For $s \in S_{s_i-i}$, the last preferred option is $(\emptyset,\emptyset)$. The penultimate option is to be all together in c_i but without s_i. Then there are all options when $s \in S_{s_i-i}$ is in c_i with s_i (except for the first one, which is located in the first place), then all other options are to be in c_i. Then all the other options go up to the first option.

$\succ_s, s \in S'$
$(c_j$, {subset of S' that contains s})
...
$to\ be\ in\ c_i\ without\ s_i$
$to\ be\ in\ c_j\ without\ s_i$
$to\ be\ in\ c_j\ with\ s_i$
$to\ be\ in\ c_i\ with\ s_i$
$(\emptyset, \emptyset)$

Each of the $s \in S'$ is in some school c_j. Next is a description of the options from bottom to top. Not being in school is the worst option. Then there are options where it student is in c_i with s_i. Then there are options when the student will be at the current school, but with s_i. After that, all remaining options to be in c_j without s_i, and then all remaining options to be in c_i without s_i.

To numerically express these options and the options that we'll use later, we introduce several notations:

$$z_l = \sum_{k=0}^{k=q_l-2} \binom{N-2}{k}$$

z_l – the number of options to be with someone specific at school c_l.

$$x_l = \sum_{k=0}^{k=q_l-1} \binom{N-1}{k} - 1$$

x_l – the number of options to be in school c_l. We have -1 since it is forbidden to be alone at school.

$$|O_s| = 1 + \sum_{l=1}^{K} x_l$$

$|O_s|$ – the number of all options.

The number of options to be with someone from $s \in S_\emptyset$ in any but one school c_i will have the value y.

$$y = \sum_{l \neq i}^{K} [\sum_{k=0}^{k=q_l-1} \binom{N-1}{k} - \sum_{k=0}^{k=q_l-1} \binom{N-1-|S_\emptyset|}{k}]$$

Let's consider the inequalities with these expressions, which we will use:

i) $x_l \geq 2z_l$

correct, since $x_l - 2z_l = \sum_{k=0}^{k=q_l-1}\binom{N-1}{k} - 2\sum_{k=0}^{k=q_l-2}\binom{N-2}{k} - 1 =$

$$= \sum_{k=1}^{k=q_l-1}\binom{N-1}{k} - 2\sum_{k=1}^{k=q_l-2}\binom{N-2}{k} - 1 + 1 - 2 =$$

$$= \sum_{k=1}^{k=q_l-1}(\binom{N-2}{k} + \binom{N-2}{k-1}) - 2\sum_{k=1}^{k=q_l-2}\binom{N-2}{k} - 2 =$$

$$= \binom{N-2}{q_l-1} + \sum_{k=1}^{k=q_l-2}\binom{N-2}{k} + \sum_{k=1}^{k=q_l-1}\binom{N-2}{k-1} - 2\sum_{k=1}^{k=q_l-2}\binom{N-2}{k} - 2 =$$

$$= \binom{N-2}{q_l-1} + \sum_{k=1}^{k=q_l-2}\binom{N-2}{k} + \sum_{k=0}^{k=q_l-2}\binom{N-2}{k} - 2\sum_{k=1}^{k=q_l-2}\binom{N-2}{k} - 2 =$$

$$= \binom{N-2}{q_l-1} - 1 \geq 0$$

ii) $y \geq z_l$

Consider the case when y is minimal, that is $|S_\emptyset| = 1$:

$$y - z_l = y = \sum_{l\neq i}^{K}[\sum_{k=0}^{k=q_l-1}\binom{N-1}{k} - \sum_{k=0}^{k=q_l-1}\binom{N-1-1}{k}] - \sum_{k=0}^{k=q_j-2}\binom{N-2}{k} =$$

$=\sum_{l\neq i,j}^{K}[\sum_{k=0}^{k=q_l-1}\binom{N-1}{k} - \sum_{k=0}^{k=q_l-1}\binom{N-1-1}{k}]+\sum_{k=0}^{k=q_j-1}\binom{N-1}{k} -$

$$- \sum_{k=0}^{k=q_j-1}\binom{N-2}{k} - \sum_{k=0}^{k=q_j-2}\binom{N-2}{k} = \sum_{l\neq i,j}^{K}[\sum_{k=0}^{k=q_l-1}\binom{N-1}{k} - \sum_{k=0}^{k=q_l-1}\binom{N-2}{k}] +$$

$$+ \sum_{k=0}^{k=q_j-1}\binom{N-2}{k} + \sum_{k=0}^{k=q_j-1}\binom{N-2}{k-1} - \sum_{k=0}^{k=q_j-1}\binom{N-2}{k} - \sum_{k=0}^{k=q_j-2}\binom{N-2}{k} =$$

$$= \sum_{l\neq i,j}^{K}[\sum_{k=0}^{k=q_l-1}\binom{N-1}{k} - \sum_{k=0}^{k=q_l-1}\binom{N-2}{k}] + \sum_{k=1}^{k=q_j-1}\binom{N-2}{k-1} - \sum_{k=0}^{k=q_j-2}\binom{N-2}{k} =$$

$$= \sum_{l\neq i,j}^{K}[\sum_{k=0}^{k=q_l-1}\binom{N-1}{k} - \sum_{k=0}^{k=q_l-1}\binom{N-2}{k}] \geq 0$$

Suppose for $\alpha \in \bar{A}(\succ_{s_i}, \succ^*_{-s_i})$ we have $\alpha(s_i) = o$ and $r_{s_i}(o) = t$ then $N * \bar{r}(\alpha) = N - 1 + t$. Now let us show that for all $\alpha' \neq \alpha$ we have $\overline{r}(\alpha') \geq \overline{r}(\alpha)$.

a1) Suppose we have α' in which $\alpha'(s) = o' = (\emptyset,\emptyset)$. Suppose $r_s(o') = 2$ which is the biggest rank for this option. Then for $s \in S_{s_i-i}$ we have option with the rank $|O_s| - 1$ because $s \in S_{s_i-i}$ is not with s_i in the school c_i anymore. $N * \overline{r}(\alpha') = N - 2 + 2 + |O_s| - 1 \geq N * \overline{r}(\alpha)$.

a2) Suppose we have α' in which anyone of other student gets the option $(\emptyset,\emptyset)$ in this case even if all other students get the option with rank 1 we have $N * \overline{r}(\alpha') = N - 1 + |O_s| \geq N * \overline{r}(\alpha)$.

a3) Suppose we have α' in which s_i moves from c_i to some c_j, which gives him rank one, then we get the following expression for the numerator of the average rank of α':

$$|S_{s_i-i}| * (|O_s| - x_i - 1 + 1) + (q_{c_j} - 1) * \left(|O_s| - z_j - z_i - 1\right) + |O_s| - x_i - x_j + N - |S_{s_i-i}| - (q_{c_j} - 1) - 1$$

Subtract from it the numerator of the average rank equals $N - 1 + t$ with $t = |O_s|$ in assumption $|S_{s_i - i}| = 1$ (the case when this value is the minimum regardless of N):

$$(|O_s| - x_i) + \left(q_{c_j} - 1\right) * \left(|O_s| - z_j - z_i - 1\right) + |O_s| - x_i - x_j + N - 1 - \left(q_{c_j} - 1\right) - 1 - (|O_s| + N - 1) = (|O_s| - x_i) - x_j - 1 + (q_{c_j} - 1) * \left(|O_s| - z_j - z_i - 2\right) - x_i$$

The expression is an increasing function of q_{c_j}, so it is enough for us to consider the case of $q_{c_j} = 2$.

$$(|O_s| - x_i) - x_j - 1 + \left(|O_s| - z_j - z_i - 2\right) - x_i = \left[|O_s| - x_i - x_j - 1\right] + [|O_s| - z_j - z_i - 2 - x_i]$$

The first expression in square brackets is always greater than or equal to zero. Consider the second expression.

$$|O_s| - z_j - z_i - 2 - x_i = 1 + \sum_{l \neq i,j}^{K} x_l + x_j - z_j + x_i - z_i - 2 - x_i = \sum_{l \neq i,j}^{K} x_l + \binom{N-2}{q_j - 1} + z_j - z_i - 1 = \sum_{l \neq i,j}^{K} x_l + \binom{N-2}{q_j - 1} - 1 \geq 0$$

So $\overline{r}(\alpha') \geq \overline{r}(\alpha)$.

a4) Suppose we have α' in which s_i stays in c_i, and someone from c_j swaps places with a student in c_i, and this placement gives s_i rank one, then we get the following expression for the numerator of the average rank of α':

$$\left(|S_{s_i - i}| - 1\right) * (|O_s| - z_i - 1) + (q_{c_j} - 1) * \left(|O_s| - x_j - z_i\right) + |O_s| - z_i - 1 + N - (|S_{s_i - i}| - 1) - (q_{c_j} - 1) - 1$$

Subtract from it the numerator of the average rank equals $N - 1 + t$ with $t = |O_s|$ in assumption $|S_{s_i - i}| = 1$ (the case when this value is the minimum regardless of N):

$$(q_{c_j} - 1) * \left(|O_s| - x_j - z_i\right) + |O_s| - z_i - 1 + N - (q_{c_j} - 1) - 1 - (|O_s| - 1 + N - 1) = (q_{c_j} - 1) * \left(|O_s| - x_j - z_i - 1\right) - z_i \geq 0$$

So $\overline{r}(\alpha') \geq \overline{r}(\alpha)$.

All other placements will affect even more agents, which means they will lead to an even greater increase in the average rank.

b) <u>Suppose we want to give student s_i option</u> $o = (c_i, \{s_i, S_{s_i - i}, \emptyset\})$.

Let us consider the preferences $\succ_{-s_i^*}$ of the following type: there are $|S_\emptyset| < q_{c_i} - 2$ students for whom not being in school (option $(\emptyset,\emptyset)$) is in the second place and to be all together in c_i without anyone else is the best option. Let us denote the set of these students $S_\emptyset$. For the rest, $S' = S/(S_{s_i} \cup S_\emptyset)$ we have preference profiles such that they occupy all available places in the remaining $K - 1$ schools and this placement has the first rank for each of them. For $S_{s_i - i}$ students, the option when they are all together with s_i in school c_i and there also has a rank of 1. Consequently, we have three sets of students: $S', S_\emptyset$, $S_{s_i - i}$. Below is a schematic description of their preferences.

$s \in S_{s_i-i}, \succ_s$
$(c_i, \{s_i, S_{s_i-i}, \emptyset\})$
$to\ be\ with\ s \in S_\emptyset\ not\ in\ c_i$
…
$to\ be\ in\ c_i\ without\ s_i$
$to\ be\ in\ c_i\ with\ s_i$
$(c_i, \{S_{s_i-i}, \emptyset\})$
$(\emptyset, \emptyset)$

For $s \in S_{s_i-i}$, the least preferred option is $(\emptyset,\emptyset)$. The penultimate option is to be all together in c_i but without s_i. Following are all options when $s \in S_{s_i-i}$ is in c_i with s_i, except for the first option, then all other options are in c_i. The best option is $(c_i, \{s_i, S_{s_i-i}, \emptyset\})$. There are options below the first one to be with $s \in S_\emptyset$ but not in c_i. Lastly, the remaining options are located in between.

$s \in S_\emptyset, \succ_s$
$(c_i, \{S_\emptyset\})$
$(\emptyset, \emptyset)$
…
$to\ be\ in\ c_i\ without\ s_i$
$to\ be\ in\ c_i\ with\ s_i$

For $s \in S_\emptyset$, to be all together in c_i without anyone else is the best option. For this students not being in school (option $(\emptyset,\emptyset)$) is in the second place. In the last place are all the options when a student from this set will be in c_i with s_i, then it will be more acceptable to be in c_i without s_i, and then all the other options.

$s \in S', \succ_s$
$(c_j, \{\text{subset of } S' \text{ that contains } s\})$
…
$to\ be\ in\ c_i\ without\ s_i$
$to\ be\ in\ c_j\ without\ s_i$
$to\ be\ in\ c_i\ with\ s_i$
$to\ be\ in\ c_j\ with\ s_i$
$(\emptyset, \emptyset)$

Each of the $s \in S'$ is in some school c_j. Not being in school is the worst option. Then there are options when the student will be at the current school, but with s_i. Then the options where it is in c_i with s_i, after which all the remaining options are in c_j, and then all the remaining options are in c_i.
Suppose for $\alpha \in \bar{A}(\succ_{s_i}, \succ^*_{-s_i})$ we have $\alpha(s_i) = o$ and $r_{s_i}(o) = t$ then $N * \bar{r}(\alpha) = N - 1 + |S_\emptyset| + t$. Now let us show that for all $\alpha' \neq \alpha$ we have $\bar{r}(\alpha') \geq \bar{r}(\alpha)$.
b1) Suppose we have α' in which $\alpha'(s) = o' = (\emptyset,\emptyset)$. Suppose $r_s(o') = 2$ which is the biggest rank for this option. Then for $s \in S_{s_i-i}$ we have option with the rank $|O_s| - 1$ because $s \in S_{s_i-i}$ is not with s_i in the school c_i anymore. $N * \bar{r}(\alpha') = N - 2 + 2 + |O_s| - 1 + |S_\emptyset| \geq N * \bar{r}(\alpha)$.
b2) Suppose we have α' in which we move $s \in S_\emptyset$ in the place which belongs $s \in S_{s_i-i}$ in

c_i, this student $s \in S_{s_i - i}$ will not attend school and will receive the rank $|O_s|$, so $N * \overline{r}(\alpha') \geq N - 1 + |O_s| + |S_\emptyset| \geq N * \overline{r}(\alpha)$.

b3) Suppose we have α' in which we move $s \in S_\emptyset$ to an empty space in c_i, and this option is the best option for s_i, then we get the following expression for the numerator of the average rank:

$$|S_{s_i-i}| * (|O_s| - 1 - z_i) + |O_s| - z_i + 1 + N - |S_{s_i-i}| - 1 + |S_\emptyset| - 1$$

Subtract from it the numerator of the average rank of the initial placement in assumption $|S_{s_i-i}| = 1$ (the case when this value is the minimum regardless of N):

$$(|O_s| - 1 - z_i) + |O_s| - z_i + 1 + N - 1 - 1 + |S_\emptyset| - 1 - (|O_s| + N - 1 + |S_\emptyset|) = |O_s| - 2z_i$$
$$= 1 + \sum_{l=1}^{K} x_l - 2z_i = 1 + \sum_{l \neq i}^{K} x_l + x_i - 2z_i \geq 0$$

So $\overline{r}(\alpha') \geq \overline{r}(\alpha)$.

b4) Suppose we have α' in which we move s_i to some c_j which will give him rank one, and some $s \in S'$ on empty place in c_i, then we get the following expression for the numerator of the average rank:

$$|S_{s_i-i}| * (|O_s| - x_i) + (q_{c_j} - 1) * \left(|O_s| - z_j + 1 - 1\right) + |O_s| - x_i - x_j + 1 - 1 + N - |S_{s_i-i}|$$
$$- (q_{c_j} - 1) - 1 + |S_\emptyset|$$

Subtract from it the numerator of the average rank of the initial placement in assumption $|S_{s_i-i}| = 1$ (the case when this value is the minimum regardless of N):

$$(|O_s| - x_i) + \left(q_{c_j} - 1\right) * (|O_s| - z_j) + |O_s| - x_i - x_j + N - 1 - \left(q_{c_j} - 1\right) - 1 + |S_\emptyset|$$
$$- (|O_s| + N - 1 + |S_\emptyset|) =$$
$$= \left[|O_s| - x_i - x_j - 1\right] + \left[\left(q_{c_j} - 1\right)\left(|O_s| - z_j - 1\right) - x_i\right] - 1 > 0$$

So $\overline{r}(\alpha') \geq \overline{r}(\alpha)$.

b5) Suppose we have α' in which $s \in S'$ is moved to an empty space in c_i and s_i gets rank one, then we get the following expression for the numerator of the average rank:

$$|S_{s_i-i}| * (|O_s| - z_i - 1 + 1) + (q_{c_j} - 1) * \left(|O_s| - x_j - z_i + 1\right) + |O_s| - z_i - z_j + N - |S_{s_i-i}|$$
$$- (q_{c_j} - 1) - 1 + |S_\emptyset|$$

Subtract from it the numerator of the average rank of the initial placement in assumption $|S_{s_i-i}| = 1$ (the case when this value is the minimum regardless of N):

$$(|O_s| - z_i) + \left(q_{c_j} - 1\right) * \left(|O_s| - x_j - z_i + 1\right) + |O_s| - z_i - z_j + 1 + N - 1 - \left(q_{c_j} - 1\right) - 1$$
$$+ |S_\emptyset| - (|O_s| + N - 1 + |S_\emptyset|)$$
$$= \left[|O_s| - z_i - z_j - 1\right] + \left[\left(q_{c_j} - 1\right)\left(|O_s| - x_j - z_i - 1\right) - z_i\right] - 1 \geq 0$$

So $\overline{r}(\alpha') \geq \overline{r}(\alpha)$.

b6) If s_i remains in c_i, but someone from S_{s_i-i} is replaced by $s \in S'$ from some c_j and s_i gets rank one, then we get the following expression for the numerator of the middle rank:

$$(|S_{s_i-i}| - 1) * (|O_s| - z_i - 1 + 1) + \left(q_{c_j} - 1\right) * \left(|O_s| - x_j - z_i + 1\right) + |O_s| - z_i - z_j + y + 1$$
$$+ N - |S_{s_i-i}| + 1 - (q_{c_j} - 1) - 1 + |S_\emptyset|$$

Subtract from it the numerator of the average rank of the initial placement in assumption $|S_{s_i-i}| = 1$ (the case when this value is the minimum regardless of N):

$$\left(q_{c_j} - 1\right) * \left(|O_s| - x_j - z_i + 1\right) + |O_s| - z_i - z_j + 1 + y + N - 1 + 1 - \left(q_{c_j} - 1\right) - 1 + |S_\emptyset|$$
$$- (|O_s| + N - 1 + |S_\emptyset|) = \left(q_{c_j} - 1\right) * \left(|O_s| - x_j - z_i\right) - z_i - z_j + y + 2$$
$$\geq \left[\left(q_{c_j} - 1\right) * \left(|O_s| - x_j - 2z_i\right)\right] + \left[y - z_j\right] + 1 > 0$$

So $\bar{r}(\alpha') > \bar{r}(\alpha)$.
All other placements will affect even more agents, which means they will lead to an even greater increase in the average rank.

c) Suppose we want to give student s_i option $o = (\emptyset, \emptyset)$.

Let us consider the preferences $\succ_{-s_i^*}$ of the following type: all students except s_i are in schools. At the same time, there is an empty place in one of the schools. The students who are in this school c_i form the set S_{-i}. The rest of the students, with the exception of s_i, are in the set S'. This placement has a rank of 1 for each student except s_i.
The descriptions of students' preferences from the two remaining sets are schematically presented below.

$s \in S_{-i}, \succ_s$
$(c_i, \{S_{-i}\})$
...
$to\ be\ in\ c_i\ without\ s_i$
$to\ be\ in\ c_i\ with\ s_i$
$(c_i, \{S_{-i}, s_i\})$
$(\emptyset, \emptyset)$

For $s \in S_{-i}$, the last preferred option is (∅,∅). The penultimate option is to be all together in c_i and with s_i. Then there are all options when $s \in S_{-i}$ is in c_i with s_i. Then all other options should be in c_i (with the exception of the first one, which is located in the first place). Then all the other options go up to the first option.

$s \in S', \succ_s$
$(c_j, \{\text{subset of } S' \text{ that contains } s\})$
...
$to\ be\ in\ c_i$
$to\ be\ in\ c_j\ without\ s_i$
$to\ be\ in\ c_j\ with\ s_i$
$(\emptyset, \emptyset)$

Each of the $s \in S'$ is in some school c_j. Not being in school is the worst option. Then there are options where it is in c_j with s_i. Then there are options when the student will be at the current school, but without s_i. After that, all remaining options should be in c_j, and then all remaining options should be in c_i.
There are only two possibilities for s_i to change his position.
c1) Suppose we have α' in which s_i can only go to c_i for an empty seat, otherwise, some of the other students will lose the opportunity to be in school, and this is the least desirable option. If s_i switch to an empty seat, there will be at least two students who will receive the penultimate option, because it is forbidden to be alone in the school. So $\bar{r}(\alpha') > \bar{r}(\alpha)$
c2) Suppose we have α' in which s_i can only switch to c_j by sending someone to c_i to an empty seat, otherwise they will not be in school, which is the worst option. If the capacity $q_{c_i} = m$, one seat is occupied by s_i, then there are $\binom{m}{m-1} = m$ options when s_i is in c_j at full occupancy. Since for any option, the student must be in this school himself, we can

place these m among the penultimate places for m students so that they do not repeat, and when one of the students is removed, there is guaranteed to be one student among the remaining ones who has this option with s_i will be in the penultimate place. Since at the same time, students from c_i lose their rank, $\overline{r}(\alpha') > \overline{r}(\alpha)$.
All other placements will affect even more agents, which means they will lead to an even greater increase in the average rank. □

Appendix B

Theorem 1: *RM algorithm (mechanism) with full support is not obviously manipulable for the case of the school choice problem with peers.*

Intuitively, while students may alter reported preferences to influence the peer composition, such manipulations never yield unambiguously better outcomes under all possible configurations of others' preferences

Proof of the Theorem 1.

For the first part of the definition:

Let's fix some $\succ_s$, we use Lemma 1:

For any $s \in S$ with $\succ_s \in P_s$ and $o \in O_s$ there exists $\succ^*_{-s} \in P_{-s}$ such that $\alpha \in \bar{A}(\succ_s, \succ^*_{-s})$ with $\alpha(s) = o$.

Suppose for $\succ^*_{-s}$ we put $\mu = \varphi^{\mathrm{RM}}(\succ_s, \succ^*_{-s})$ and there exists $\alpha \in \bar{A}(\succ_s, \succ^*_{-s})$ with $\alpha(s) = o$ s.t. $r_s(\alpha(s)) = |O|$ then $r_s(\alpha(s)) = \max_{\succ_{-s}} \overline{\rho_s}(\varphi^{RM}(\succ_s, \succ_{-s}))$ with $\mu(\alpha) > 0$.

Now suppose under some manipulation $\succ'_s$ we have $\varphi^{RM}(\succ'_s, \succ^*_{-s}) = \mu'$ and there exists $\alpha^* \in \bar{A}(\succ'_s, \succ^*_{-s})$ with $\mu'(\alpha^*) > 0$. And $\overline{r'}(\alpha^*) < \overline{r'}(\alpha)$ because now α^* is an outcome of φ^{RM}, where $\overline{r'}$ - is the average rank over the fake preferences $\succ'_s$, but according to Lemma 1 there exists $\succ^{**}_{-s} \in P_{-s}$ such that $\alpha^* \in \bar{A}(\succ'_s, \succ^{**}_{-s})$ with $\alpha^*(s) = o$ and because under true $\succ_s$ we have $r_s(o) = |O|$ then $\max_{\succ_{-s}} \overline{\rho_s}(\varphi^{RM}(\succ_s, \succ_{-s})) \leq \max_{\succ_{-s}} \overline{\rho_s}(\varphi^{RM}(\succ'_s, \succ_{-s}))$.

For the second part of the definition:

it is obviously true because $\min_{\succ_{-s}} \underline{\rho_s}(\varphi^{RM}(\succ_s, \succ_{-s})) = 1 \leq \min_{\succ_{-s}} \underline{\rho_s}(\varphi^{RM}(\succ'_s, \succ_{-s})) = 1.$ □

Chapter 2

Non-obvious manipulability and regret-free truth-telling for the School Choice problem

Andrei Belykh

We consider the weak strategic properties of the mechanisms for the school choice problem. We have shown that for unanimous mechanisms, the regret-free truth-telling implies non-obvious manipulability. At the same time, the property of non-obvious manipulability does not imply the regret-free truth-telling, which is shown by the example of the RM-SD mechanism.

Key Words: matching, school choice problem, non-obvious manipulability, regret-free truth-telling, strategy-proof mechanism

1 Introduction

The school choice problem (SCP) models the assignment of students to public schools based on students' preferences and schools' priorities (Abdulkadiroğlu and Sönmez, 2003). Since the seminal work of Gale and Shapley (1962), various mechanisms have been proposed to implement fair and efficient assignments in such environments. These mechanisms, such as the Deferred Acceptance (DA), Boston, and Serial Dictatorship (SD) algorithms, differ in terms of stability, efficiency, and the degree to which they incentivize truthful preference revelation (Sönmez and Ünver, 2009; Ergin and Sönmez, 2006; Pathak and Sönmez, 2008).

While the strategy-proofness property ensures that truth-telling is a dominant strategy for all agents, many school choice mechanisms fail to meet this strong requirement. In recent years, the literature has therefore turned to weaker notions of incentive compatibility that better capture boundedly rational behavior and realistic decision-making. Among these, three properties have attracted particular attention: obvious strategy-proofness (Li, 2017), and its behavioral relaxations, non-obvious manipulability (Troyan and Morrill, 2020) and regret-free truth-telling (Fernandez, 2020).

Li (2017) introduced the concept of obvious strategy-proofness (OSP) to capture mechanisms where the truth-telling strategy is not only dominant but also evidently optimal, even for agents with bounded rationality. His framework builds on the extensive-form representation of mechanisms, emphasizing the cognitive accessibility of optimal strategies. In this sense, OSP mechanisms are robust to bounded rationality and have been applied to matching and allocation problems (Pycia and Troyan, 2016; Ashlagi and Gonczarowski, 2018). However, as Li (2017) showed, many well-known mechanisms, such as the student-proposing Deferred Acceptance, are strategy-proof but not obviously strategy-proof.

Troyan and Morrill (2020) relaxed Li's notion and introduced the concept of non-obvious manipulability (NOM), which allows for mechanisms that are manipulable but in which any profitable deviation is not obvious to the agent. In this sense, a mechanism may fail strategy-proofness while remaining robust to limited rationality, since agents cannot clearly identify a deviation that is guaranteed to improve their outcome. This property has been used to analyze manipulable school choice mechanisms, such as the Boston mechanism and the school-proposing Deferred Acceptance (SP-DA) algorithm, and more recently, the Rank-Minimizing (RM) mechanism (Troyan, 2024). By distinguishing between obvious and non-obvious incentives to misreport, NOM provides a refined behavioral foundation for evaluating mechanisms that are not fully strategy-proof but may still encourage truthful behavior in practice.

In parallel, Fernandez (2020) proposed the concept of regret-free truth-telling (RFTT), focusing on agents' *ex post* evaluation of their reported preferences. A mechanism satisfies RFTT if, after observing the outcome and others' possible reports, an agent never regrets having reported truthfully. This property, inspired by ex post rationalizability and regret-based learning (Bourgeois-Gironde, 2010), was later explored in applied settings by Chen and Möller (2024), who showed that the efficiency-adjusted DA mechanism satisfies RFTT under consent-based school choice.

The concept of regret-free truth-telling (Fernandez, 2020) aligns with a growing literature that connects mechanism design with regret-based learning (Mandler, 2006; Fudenberg and Levine, 1998; Hart and Mas-Colell, 2000). In this view, agents evaluate their reporting strategies ex post, comparing realized outcomes with counterfactual ones. Mechanisms that are regret-free truth-telling therefore ensure not only cognitive but also dynamic stability of truthful reporting, even in repeated or feedback-rich environments.

Recent work in behavioral mechanism design emphasizes the need to evaluate mechanisms not only for their formal incentive properties but also for their cognitive accessibility (Li, 2017; Pycia and Troyan, 2016; Ashlagi and Gonczarowski, 2018). Experimental and theoretical studies suggest that agents frequently fail to identify profitable deviations in complex mechanisms (Kawagoe and Kojima, 2020; Ding and Schotter, 2021), underscoring the importance of behavioral notions such as non-obvious manipulability and regret-free truth-telling.

From a policy perspective, understanding weak incentive properties is crucial because real-world school choice systems often trade off efficiency against full strategy-proofness (Pathak and Sönmez, 2008; Ortega and Klein, 2023). Empirical studies show that even when strategy-proof mechanisms are implemented, agents may deviate due to bounded rationality or peer effects (Featherstone and Niederle, 2016; Dur and Wiseman, 2019). Behavioral properties such as non-obvious manipulability and regret-free truth-telling thus provide a more realistic lens through which to assess the robustness of allocation mechanisms.

Despite their conceptual similarity, the relationship between non-obvious manipulability and regret-free truth-telling has not been formally studied. Both properties aim to describe mechanisms that are resilient to bounded rationality and psychological deviations from full strategy-proofness, yet they rely on distinct informational assumptions: NOM focuses on agents' ability to anticipate payoffs under uncertainty, while RFTT relies on agents' *ex post* reasoning about alternative reports.

This paper establishes a formal connection between these two behavioral properties. We show that for unanimous mechanisms, a minimal regularity condition satisfied by most practical school choice algorithms, regret-free truth-telling implies non-obvious manipulability. Intuitively, if agents never regret truthful reporting, then no manipulation can be *obviously* profitable ex ante. However, the converse does not hold: non-obvious manipulability does not guarantee regret-free truth-telling. We demonstrate this by providing a counterexample based on the Rank-Minimizing Serial Dictatorship (RM–SD) mechanism, which is known to be non-obviously manipulable (Troyan, 2024) but fails the RFTT property.

Our results contribute to the literature on behavioral mechanism design by unifying two strands of research: (i) the study of cognitive robustness in matching mechanisms (Li, 2017; Troyan and Morrill, 2020; Troyan, 2024), and (ii) the analysis of ex post rationality and regret (Fernandez, 2020; Chen and Möller, 2024). Conceptually, the paper clarifies that regret-free truth-telling is a strictly stronger property than non-obvious manipulability, providing a hierarchy of incentive notions below strategy-proofness. From a practical perspective, this insight informs the design of school choice procedures that are robust to cognitive limitations and behavioral biases without requiring full strategy-proofness (Ortega and Klein, 2023; Pathak and Sönmez, 2013).

The remainder of the paper is organized as follows. Section 2 introduces the formal model of the school choice problem. Section 3 defines the strategic and behavioral properties under consideration. Section 4 establishes the main results connecting RFTT and NOM and provides the illustrative example of the RM–SD mechanism. Section 5 concludes.

2 The Model

We study the following formal problem.(Abdulkadiroglu and Sonmez, 2003). Let C be a set of schools and S be a set of students (with $|S| = N, |C| = K$). Each school c has a strict priorities $\succ_c$ over 2^S. Each student s has a strict preference ordering $\succ_s$ over possible options $O_s \equiv C \cup \{\emptyset\}$. Each school has a capacity q_m, $\sum_{m=1}^{K} q_m \geq N,\ \forall m \in R,\ q = (q_1, \dots, q_K)$. The $\succ_S = (\succ_{s_1}, \dots, \succ_{s_N})$ is preference profile and $\succ_C = (\succ_{c_1}, \dots, \succ_{c_K})$ is the priority profile. From now we refer to the School Choice Problem (or economy) as tuple $\langle S, C, \succ_S, \succ_C, q \rangle$.

Definition 1: A matching $\mu: C \cup S \to \{(2^S \times C)\} \cup \{\emptyset\}$ is a mapping that satisfies:

- $\mu(c) \in 2^S\ \forall c \in C$ and $|\mu(c)| \leq q_c$
- $\mu(s) \in C \cup \{\emptyset\} \forall s \in S$ and $|\mu(s)| = 1$
- $s \in \mu(c) \leftrightarrow \mu(s) = c$

M is the set of possible matchings μ.

A matching specifies where each student is assigned and how many students each school accepts.

Definition 2: A mechanism (an algorithm) φ is a function: $\succ_S \times \succ_C \to M$.

It should be noted that since some of the mechanisms do not use the priorities of schools, in this case the mechanism will be determined as follows: $\succ_S \to \mathrm{M}$. We examine properties related to the strategic interaction of agents, with a minimal required property in mind, which we refer to as unanimity. It should be noted that schools are not strategic agents so the school priorities and capacities are public information and are known to all of the students.

Economically, the model describes a centralized assignment process: students submit applications to schools, schools have limited capacity and fixed priority structures, and the mechanism acts as a planner that determines the final allocation. This formulation encompasses all major mechanisms studied in the school choice literature, including the student-proposing Deferred Acceptance, the Boston mechanism, Serial Dictatorship, and the SD-Rank-Minimizing mechanism examined later in the paper.

3 Properties

We will start by defining manipulation and strong property - strategy-proofness. We consider only profitable manipulations and define the utility function in the same way as in the paper (Troyan and Morrill, 2020).

Definition 3: Report $\succ_i' \in P_i$ is a profitable manipulation of a mechanism φ for student i of type $\succ_i$ if there exists some $\succ_{-i}$ such that

$$u_i(\varphi(\succ_i', \succ_{-i})) > u_i(\varphi(\succ_i, \succ_{-i}))$$

Definition 4: We say that the mechanism is strategy-proof if there are no profitable manipulations.

We define regret-free truth-telling and non-obvious manipulability in a manner consistent with the original papers (Fernandez, 2020; Troyan and Morrill, 2020), while also incorporating additional considerations specific to the school choice problem. First, we outline the definition of regret-free truth-telling for a mechanism.

Given a mechanism φ, and a school choice problem $\langle S, C, \succ_S, \succ_C, q\rangle$, suppose student i reports $\succ_i'$, observes matching μ and schools priority profile $\succ_C$. Then i's inference set, denoted $I(\mu, \succ_i', \varphi) = \{(\succ_{-i}, \succ_C) \in \succ_S \times \succ_C \colon \varphi(\succ_i', \succ_{-i}, \succ_C) = \mu\}$ identifies the preference profiles of other agents that are consistent with the observed matching, given her report, and the known rules of the mechanism.

N.B. For brevity, and given that schools are not strategic agents, we will henceforth use $(\succ_i, \succ_{-i})$ to represent the full preference profile, omitting the school priorities $\succ_C$. Additionally, we will use $\succ_{-i}$ to denote the pair $(\succ_{-i}, \succ_C)$.

Definition 5: Given a mechanism φ and an observed matching μ, agent i regrets reporting $\succ_i'$ if there is an alternative report $\succ_i''$ such that

1) for any $\succ_{-i} \in I(\mu, \succ_i', \varphi)$ it holds that $\varphi(\succ_i'', \succ_{-i}) \succsim_i \mu$ and[1]
2) for some $\tilde{\succ}_{-i} \in I(\mu, \succ_i', \varphi)$ it holds that $\varphi(\succ_i'', \tilde{\succ}_{-i}) \succ_i \mu$

Definition 6: Given a mechanism φ, a report $\succ_i'$ is regret-free for student i, if it is not possible for i to regret about $\succ_i'$.

Definition 7: A mechanism φ is regret-free truth-telling if for any matching, and any agent, truth-telling is regret-free.

The property of regret-free truth-telling (RFTT) captures a weak behavioral form of incentive compatibility based on agents' *ex post* evaluation of their reports. In a direct revelation mechanism, each student submits a ranking of schools before the mechanism determines an assignment. After the outcome is realized, a student can compare the school she actually receives when reporting truthfully with the schools she could have received under alternative reports, given that other students' reports remain fixed. A mechanism satisfies RFTT if, for every student and every possible configuration of others' reports, the student never finds that an alternative report would have resulted in a strictly preferred assignment once the outcome is known. In other words, after observing the

[1] That is, for any $c, c_1 \in C \cup \{\emptyset\}$, $c \succsim_i c_1$ if and only if $c \succ_i c_1$ or $c = c_1$

final matching, truth-telling generates no *ex post* regret because the student cannot identify any misreport that would have produced a better result for her.

Conceptually, RFTT differs from standard strategy-proofness, which requires that truthful reporting dominate all deviations *ex ante*, before the mechanism operates. RFTT is weaker and more behavioral in nature, as it focuses on how agents reason *after* outcomes are realized rather than on how they form optimal strategies beforehand. This distinction makes RFTT especially suitable for environments with boundedly rational participants who may fail to anticipate all contingencies but are still able to recognize, after the outcome is known, whether truth-telling was a mistake.

Next, we define what it means for a mechanism to be non-obviously manipulable. Our definition follows the framework of Troyan and Morrill (2020) but extends it in one important respect. In particular, we include school priority profiles in the set over which the minimum and maximum are taken, even though schools are not strategic agents in the school choice problem. This modification ensures that the definition captures all relevant sources of variation in outcomes and provides a consistent comparison across mechanisms that differ in their treatment of priorities.

Definition 8: Mechanism φ is not obviously manipulable if, for every student i, for any profitable manipulation $\succ_i'$ the following are true:

3) $\min\limits_{\succ_{-i}} u_i(\varphi(\succ_i', \succ_{-i})) \leq \min\limits_{\succ_{-i}} u_i(\varphi(\succ_i, \succ_{-i}))$
4) $\max\limits_{\succ_{-i}} u_i(\varphi(\succ_i', \succ_{-i})) \leq \max\limits_{\succ_{-i}} u_i(\varphi(\succ_i, \succ_{-i}))$

If either 1) or 2) does not hold for some manipulation $\succ_i'$ then $\succ_i'$ is said to be an obvious manipulation for student i of type $\succ_i$, and mechanism φ is said to be obviously manipulable (OM).

The property of non-obvious manipulability (NOM) describes a behavioral form of robustness that allows for possible manipulations but requires that any profitable deviation be cognitively non-trivial. Under a direct revelation mechanism, a student may have an incentive to misreport her preferences if this can yield a more preferred school for some configuration of others' reports or priority orders. However, a mechanism is said to be non-obviously manipulable if, for every student, no single misreport guarantees a strictly better outcome across all relevant states of the world. In other words, while a manipulation may exist, the potential gain is not *obvious* because the student cannot identify a deviation that is uniformly profitable once the range of possible environments is taken into account.

Conceptually, NOM lies between strategy-proofness and unrestricted manipulability. It relaxes the full dominance requirement of strategy-proofness but maintains behavioral stability by ruling out manipulations that are clearly and universally beneficial. This notion captures the idea that boundedly rational agents may not exploit mechanisms unless the advantage of deviating is transparent and unambiguous. By focusing on the cognitive accessibility of profitable deviations rather than their mere existence, NOM provides a way to evaluate mechanisms that are not fully strategy-proof yet still robust to limited strategic reasoning.

To obtain the main result, we introduce a very weak requirement of unanimity.

Definition 9: A mechanism φ is unanimous if for any $\succ_S \in P_S$ such that there exist a $\mu' \in M$ such that for any $i \in S$ and any $\mu \in M$ $\mu' \succsim_i \mu$ implies $\varphi(\succ_S) = \mu'$.

Note that whenever $\mu' \sim_i \mu$ for all $\mu, \mu' \in M$ and $i \in S$, $\mu'(i) = \mu(i)$ and then $\mu' = \mu$.

The property of unanimity requires a mechanism to respect a complete consensus among students whenever such an assignment is feasible. If all students rank the same school as their top choice and the school has enough capacity to admit them, the mechanism must assign every student to that school. This condition represents a minimal form of collective consistency and is satisfied by most standard school choice mechanisms.

Definition 10: We define $\tau(\succ_i) = c$ as a top choice for student i according to her preferences.

4 Propositions and proofs

We will demonstrate that, for the school choice problem, every unanimous regret-free truth-telling mechanism is not obviously manipulable. First, we will show that unanimity implies the second condition of non-obvious manipulability (NOM). Next, we will prove that the regret-free truth-telling (RFTT) property implies the first condition of NOM. Finally, we provide an example of a mechanism that satisfies NOM but not RFTT, thereby concluding our analysis of the relationship between these properties.

Proposition 1: For any unanimous mechanism the following is true: For any student i for all profitable manipulation $\succ_i'$:

$$\max_{\succ_{-i}} u_i(\varphi(\succ_i', \succ_{-i})) \leq \max_{\succ_{-i}} u_i(\varphi(\succ_i, \succ_{-i}))$$

Proof: First, assume that the mechanism is strategy-proof, i.e., no profitable manipulation exists. In this case, Proposition 1 holds trivially. Next, consider the case in which the mechanism is manipulable. Let us fix some i with $\succ_i$ and let us consider $\succ_S$ such that $|\{s \in S | \tau(\succ_S) = c\}| = q_c$. This profile of preferences exists because of $\sum_{m=1}^{K} q_m \geq N$. Now $\varphi(\succ_S) = \mu$ and for any $s \in S$, $\mu(\mathrm{s}) = \tau(\succ_s)$, because φ is unanimous. Then $\mu(i) = \tau(\succ_i)$ and so for any $\succ_i' \neq \succ_i$

$$\max_{\succ_{-i}} u_i(\varphi(\succ_i', \succ_{-i})) = u_i(\tau(\succ_i)) = \max_{\succ_{-i}} u_i(\varphi(\succ_i, \succ_{-i})) \ \square$$

Proposition 2: For any unanimous mechanism regret-free truth-telling implies nonobvious manipulability.

Proof: Because of Proposition 1 we need to show that for φ for any profitable manipulation $\succ_i'$:

$$\min_{\succ_{-i}} u_i(\varphi(\succ_i', \succ_{-i})) \leq \min_{\succ_{-i}} u_i(\varphi(\succ_i, \succ_{-i}))$$

First, assume that the mechanism is strategy-proof, i.e., no profitable manipulation exists. In this case, mechanism is RFTT and NOM. Now, consider the case where the mechanism is manipulable. Assume for the sake of contradiction, that for some $\succ_i$ $\exists \succ_{-i}^*$ and a profitable manipulation $\succ_i'$ such that

$$u_i\big(\varphi(\succ_i, \succ_{-i}^*)\big) < \min_{\succ_{-i}} u_i(\varphi(\succ_i', \succ_{-i})) \qquad (*)$$

Consider the notion of regret from Definition 5. Given the matching $\mu = \varphi(\succ_i, \succ_{-i}^*)$ agent i regrets reporting his true preferences $\succ_i$ over $\succ_i'$. Not only for profiles in the inference set $I(\mu, \succ_i, \varphi)$ but for every $\succ_{-i}$ it follows from (*) that $u_i\big(\varphi(\succ_i, \succ_{-i}^*)\big) < u_i\big(\varphi(\succ_i', \succ_{-i})\big)$.

Since i regrets reporting the truth, the mechanism is not RFTT.□

Proposition 3: The RM-SD mechanism is not regret-free true-telling for the case of the school choice problem.

Proposition 4: Non-obvious manipulability does not imply regret-free true-telling.

Proof: We provide an example to show that NOM does not imply RFTT.

For the school choice problem in unit capacity markets, the RM-SD mechanism is not obviously manipulable (Troyan, 2024). The RM-SD mechanism is described in Appendix A.

To prove the statement, it is enough to give an example where a student feels regret in some situation. Let us consider an example of the economy $\langle S, C, \succ_S, \succ_C, q\rangle$. $S = \{s_1, s_2, s_3\}, C = \{c_1, c_2, c_3\}$ and we ignore priorities of the school $\succ_C$. The school capacities are $\forall\, m \in \{1,2,3\}$ $q_{c_m} = 1$. The ordering of the agents is the following $f(1) = s_2, f(2) = s_3, f(3) = s_1$.

Suppose $\varphi^{RM-SD}(\succ_S) = \mu,$ $\mu = (\{c_1, s_2\}, \{c_2, s_3\}\, \{c_3, s_1\})$ and $r_{s_1}(\mu) = 3$. The student s_1 observes μ. $I(\mu, \succ_{s_1}, \varphi^{RM-SD}) = \{\succ_{-s_1.1}, \succ_{-s_1.2}, \succ_{-s_1.3}, \succ_{-s_1.4}, \succ_{-s_1.5}, \succ_{-s_1.6}\}$ Where

$(\succ_{s_1}, \succ_{-s_1.1})$:

$\succ_{s_1}$	$\succ_{s_2}$	$\succ_{s_3}$
c_1	c_1	c_1
c_2	c_2	c_2
c_3	c_3	c_3

$(\succ_{s_1}, \succ_{-s_1.2})$:

$\succ_{s_1}$	$\succ_{s_2}$	$\succ_{s_3}$
c_1	c_1	c_2
c_2	c_2	c_1
c_3	c_3	c_3

$(\succ_{s_1}, \succ_{-s_1.3})$:

$\succ_{s_1}$	$\succ_{s_2}$	$\succ_{s_3}$
c_1	c_1	c_2
c_2	c_2	c_3
c_3	c_3	c_1

$(\succ_{s_1}, \succ_{-s_1.4})$:

$\succ_{s_1}$	$\succ_{s_2}$	$\succ_{s_3}$
c_1	c_1	c_2
c_2	c_3	c_3
c_3	c_2	c_1

$(\succ_{s_1}, \succ_{-s_1.5})$:

$\succ_{s_1}$	$\succ_{s_2}$	$\succ_{s_3}$
c_1	c_1	c_2
c_2	c_3	c_1
c_3	c_2	c_3

$(\succ_{s_1}, \succ_{-s_1.6})$:

$\succ_{s_1}$	$\succ_{s_2}$	$\succ_{s_3}$
c_1	c_1	c_1
c_2	c_3	c_2
c_3	c_2	c_3

Suppose the fake preferences $\succ'_{s_1}$ of the student s_1 are the following:

$\succ'_{s_1}$
c_2
c_1
c_3

For this $\succ'_{s_1}$ the following are true:

$$\varphi^{RM-SD}(\succ'_{s_1}, \succ_{-s_1.1}) \succ_{s_1} \mu$$

$$\varphi^{RM-SD}(\succ'_{s_1}, \succ_{-s_1.2}) \succsim_{s_1} \mu$$

$$\varphi^{RM-SD}(\succ'_{s_1}, \succ_{-s_1.3}) \succ_{s_1} \mu$$

$$\varphi^{RM-SD}(\succ'_{s_1}, \succ_{-s_1.4}) \succ_{s_1} \mu$$

$$\varphi^{RM-SD}(\succ'_{s_1}, \succ_{-s_1.5}) \succsim_{s_1} \mu$$

$$\varphi^{RM-SD}(\succ'_{s_1}, \succ_{-s_1.6}) \succsim_{s_1} \mu$$

From this follows that student s_1 regret reporting $\succ_{s_1}$. □

5 Discussion and conclusion

This paper examines two behavioral notions of incentive compatibility: non-obvious manipulability (NOM) and regret-free truth-telling (RFTT) in the context of the school choice problem. Both properties relax the strong informational and cognitive assumptions that underlie classical strategy-proofness and provide a more realistic foundation for evaluating mechanisms when agents have limited reasoning abilities. The main result shows that, for mechanisms satisfying unanimity, RFTT implies NOM. This finding establishes a formal hierarchy of behavioral incentive properties and clarifies that regret-free truth-telling is a strictly stronger notion than non-obvious manipulability. The hierarchy with examples of mechanisms is shown in Figure 1.

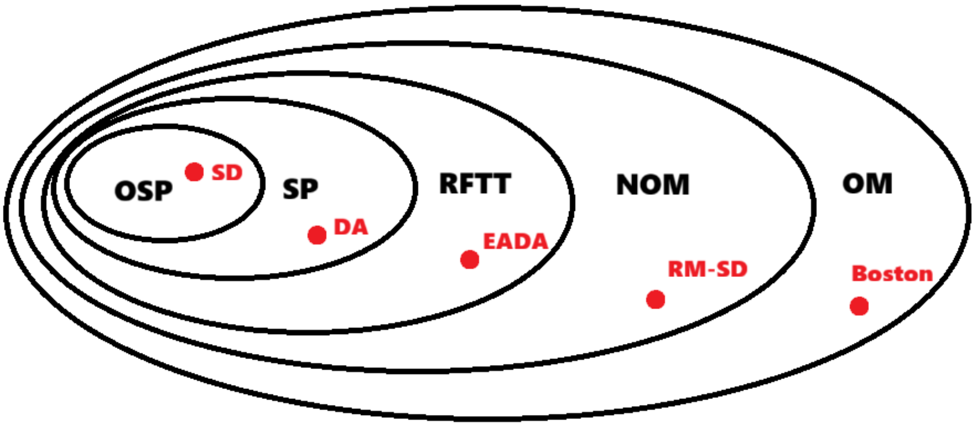

Figure 1. Hierarchy of behavioral incentive properties

The figure illustrates the logical relationship among five behavioral notions of incentive compatibility. Obviously, strategy-proof mechanisms (OSP) form the most restrictive class, followed by strategy-proof (SP), regret-free truth-telling (RFTT), and non-obvious manipulability (NOM). Obvious manipulability (OM) is the broadest class. Each red dot represents a canonical mechanism that satisfies the corresponding property: Serial Dictatorship (SD) for OSP, Deferred Acceptance (DA) for SP, Efficiency-Adjusted Deferred Acceptance (EADA) for RFTT, Rank-Minimizing Serial Dictatorship (RM–SD) for NOM, and the Boston mechanism for OM.

The result has several conceptual implications. First, it unifies two strands of the behavioral mechanism design literature. The analysis of cognitive robustness initiated by Li (2017) and extended by Troyan and Morrill (2020) focuses on agents' reasoning before the mechanism operates, while the regret-based approach of Fernandez (2020) and Chen and Möller (2024) emphasizes how agents evaluate their choices after the outcome is realized. By linking these frameworks, we show that mechanisms that eliminate ex post regret also eliminate obvious opportunities for profitable deviation ex ante. In this sense, RFTT provides not only a dynamic but also a cognitively stable form of truthful behavior.

Second, the connection between RFTT and NOM highlights how behavioral refinements can guide practical design without imposing the demanding constraints of full strategy-proofness. Most real-world school choice systems balance efficiency against incentives for truthful reporting. Mechanisms that satisfy RFTT or NOM allow policymakers to preserve nearly truthful participation while improving welfare relative to fully strategy-proof benchmarks such as the student-proposing Deferred Acceptance mechanism. Because these weaker properties rely on realistic assumptions about agents' reasoning, they may better predict behavior in empirical and experimental settings (Featherstone and Niederle, 2016; Rees-Jones, 2018; Kawagoe and Kojima, 2020).

Third, the results underscore the conceptual importance of unanimity as a minimal consistency condition for behavioral analysis. This assumption ensures that the comparison between mechanisms is grounded in a shared baseline of collective rationality. In practice, unanimity is satisfied by most canonical algorithms, which makes the theoretical relationship between RFTT and NOM broadly applicable to existing school choice systems.

The findings also open an avenue for future research that extends beyond the school choice framework. An important question is whether the implication from regret-free truth-telling to non-obvious manipulability can be established in more general allocation environments. Future work could examine whether a similar relationship holds in broader classes of matching and assignment problems, including markets with contracts, capacity constraints, or externalities. Such an extension would clarify whether the hierarchy between these behavioral incentive properties reflects a fundamental structural feature of mechanism design rather than one specific to the school choice problem.

Finally, the analysis suggests a broader methodological lesson. Behavioral mechanism design should not only ask whether truthful reporting is optimal in principle but also whether it is recognizably and regretlessly optimal for real participants. Properties such as NOM and RFTT provide intermediate benchmarks that connect normative theory with descriptive realism. Understanding the logical relationships among these properties, including strategy-proofness, regret-free truth-telling, and non-obvious manipulability, offers a framework for designing mechanisms that are both theoretically sound and behaviorally credible.

From a policy perspective, the relationship between RFTT and NOM has direct relevance for the design of centralized school assignment systems. In many cities, policymakers face a trade-off between transparency, efficiency, and resistance to manipulation. Mechanisms that are strategy-proof, such as the student-proposing Deferred Acceptance algorithm, often sacrifice allocative efficiency, while more efficient mechanisms, such as the Boston or rank-minimizing variants, expose participants to potential gains from misreporting. The results of this paper suggest that behavioral robustness does not necessarily require full strategy-proofness. Mechanisms that satisfy RFTT automatically prevent obvious manipulations, which ensures that even boundedly rational agents perceive truthful reporting as safe and satisfactory.

This insight supports the development of hybrid assignment systems that combine efficiency with cognitive stability. For example, implementing mechanisms that satisfy RFTT can reduce the informational burden on participants while maintaining fairness and predictability of outcomes. In settings where transparency and public understanding are critical, emphasizing RFTT may increase participation and trust in centralized allocation procedures. By translating abstract behavioral properties into actionable design principles, policymakers can create systems that are both efficient and psychologically sustainable.

In summary, this paper contributes to the growing literature on behavioral mechanism design by formally connecting cognitive and regret-based perspectives on truthful behavior. By establishing that regret-free truth-telling implies non-obvious manipulability under unanimity, it provides a unified theoretical foundation for mechanisms that are efficient, comprehensible, and robust to limited rationality. The broader conclusion is that incorporating behavioral realism into mechanism design can yield institutions that perform effectively not only in theory but also in practice.

Appendix A

Rank-Minimizing Serial Dictatorship (RM-SD) mechanism.

For a set of matchings $M^* \subseteq M$ and an agent preference $\succ_i$, let $Top_i(M^*) = \{\mu \in M^*: \mu \succsim_i \mu' \text{ for all } \mu' \in M^*\}$. The rank of option o according to i's preferences is

$$r_i(o) = |\{o' \in O: o' \succsim_i o\}|$$

Given a preference profile $\succ_S$ we denote the average rank of any matching μ (with respect to $\succ_S$) by

$$\bar{r}(\mu) = 1/N \sum_{i=1}^{N} r_i(\mu(i))$$

and let $M' = \{\mu \in M: \bar{r}(\mu) \leq \bar{r}(\mu') \ \forall \mu' \in M\}$.
We call $M'(\succ_S)$ the set of rank-minimizing matchings at $\succ_S$.
The RM-SD mechanism is defined as follows.

* Fix a bijection $f: \{1, \dots, N\} \rightarrow S$. This bijection produces an ordering of the agents, where agent (1) is the first, agent (2) is the second, etc.
* Given a reported preference profile $\succ_S$, let $M'(\succ_S)$ be the set of rank-minimizing matchings at $\succ_S$. Initialize $M_0 = M'(\succ_S)$.
* Consider agent (1) = i, and calculate the set $M_1 = Top_i(M_0)$.
* Consider agent (2) = i, and calculate the set $M_2 = Top_i(M_1)$.
* etc.

The mechanism ends with a final set $M_N = \{\mu\}$. That is M_N contains a unique matching μ. This matching μ is the final output of the RM-SD mechanism.

Chapter 3

The condition for the stability of coalitions in the reinsurance market

Andrei Belykh

This paper studies the formation of reinsurance pools in a decentralized insurance market. We model the problem as a hedonic coalition game in which each insurer's utility depends solely on the composition of its coalition, through the members' capital endowments and bankruptcy probabilities. The market structure requires each coalition to maintain aggregate reserves sufficient to cover the potential default of at least one of its members. Under mild local homogeneity in risk characteristics and strictly monotone preferences for more reliable partners, we show that the induced hedonic game admits a unique core-stable partition. The stable coalitions consist of consecutive sets of the most reliable firms in the reliability ordering, and no alternative grouping can be supported in the core. This result characterizes the endogenous structure of risk-sharing arrangements and demonstrates how solvency regulation shapes coalition stability.

Key Words: hedonic games, coalition stability, reinsurance pooling, stable partition, insurance networks, risk-sharing, core stability, coalition formation.

1 Introduction

Reinsurance pooling enables insurers to share the losses generated by low-probability, high-severity events. In some markets this is arranged through centralized public reinsurance schemes, while in others it is decentralized and relies on voluntary coalitions of private insurers. In the decentralized case, each member contributes capital to a joint reserve that absorbs losses if one of the participating insurers defaults. A key question is whether such coalitions are stable: under what conditions will each member prefer its existing partners to any alternative grouping?

We study this question in a setting where insurers differ in their financial strength and default risk, and where regulation limits the scale of liabilities relative to capital. This solvency constraint implies that a coalition must hold sufficient aggregate reserves to cover the failure of at least one member. Conditional on satisfying this constraint, an insurer prefers to pool with partners that are more financially reliable and to avoid coalitions that are unnecessarily large, since larger pools increase the exposure to the default risk of others.

These forces generate hedonic coalition preferences driven by two opposing effects: diversification benefits from sharing risk, and counterparty-risk exposure arising from less reliable partners. We show that, under mild conditions ensuring that insurers are not too heterogeneous, these preferences lead to a unique core-stable coalition structure. The stable coalitions consist of groups of insurers with similar reliability, formed at the minimal size required to satisfy the solvency constraint. No alternative grouping of firms can block the stable arrangement.

This result provides a sharp characterization of how solvency regulation shapes the endogenous structure of private reinsurance networks. When firms are sufficiently similar in terms of financial soundness, stable pooling arises naturally. When they are too heterogeneous, no stable coalition structure exists and centralized intervention may be required.

2 Literature Review

The problem of forming reinsurance pools among insurance companies can be modeled as a hedonic coalition game, where each company's utility depends solely on the composition of the coalition it joins. This modeling approach is inspired by the work of Bogomolnaia and Jackson (2002), who analyze hedonic coalition structures under various notions of stability such as individual, Nash, and core stability. Their framework directly supports the definition of stability used in our reinsurance pooling model, where firms form coalitions (pools) based on preferences over the identities and risk profiles of other participating firms. In particular, insurers tend to prefer coalitions with more solvent or low-risk members, as such partners reduce the probability of having to absorb others' defaults.

This yields a purely hedonic preference structure: the utility an insurer derives from a coalition depends not on total surplus or monetary payoff per se, but on who else is in the coalition — i.e., the distribution of capital and default probabilities within the pool. Under this framework, we can formally apply known solution concepts such as core-stability. The connection between hedonic games and reinsurance pooling thus allows us to rigorously explore the existence and structure of stable reinsurance coalitions under minimal institutional assumptions.

Lazarova et al. (2005, 2009) examine coalition formation under explicit contractual constraints on admissible deviations and distinguish between individual, contractual, and compensational stability. Their framework is relevant for our setting in that the solvency requirement acts as a coalition contract: only coalitions satisfying the reserve condition are feasible, and thus not all blocking deviations are institutionally permissible. Moreover, their assumption of myopic decision-making—agents evaluate coalitions based solely on current membership—corresponds directly to the hedonic formulation we adopt, where each insurer’s utility depends only on the characteristics of its coalition partners. As in their analysis, the structure of feasible deviations is endogenous to the institutional environment; here, regulatory capital constraints determine which coalitions can form and which deviations can sustain stability. Our model can therefore be interpreted as a special case of contract-constrained hedonic coalition formation in which feasibility is governed by solvency regulation, and where preferences are strictly monotone in partner reliability.

Dimitrov, Lazarova, and Sung (2016) study how stability in hedonic coalition formation can be achieved by modifying the underlying game, either through adjustments in preferences or through the introduction of institutional constraints that restrict the set of feasible deviations. Their notion of an implementation core characterizes coalition structures that can be made stable via minimal interventions. This perspective is directly relevant to our setting: the solvency requirement functions as an exogenous institutional constraint that endogenously limits which coalitions insurers may form. In effect, regulation “edits” the underlying hedonic game by ruling out coalitions that cannot meet the reserve condition, thereby altering the deviation structure in favor of stability. Moreover, in contrast to their general mechanism of preference perturbation, stability in our model arises endogenously from the joint effect of monotone partner-reliability preferences and the feasibility restriction. Thus, our results can be interpreted as identifying conditions under which solvency regulation itself implements a core-stable coalition structure without requiring further compensatory adjustments.

Although insurers in our model strictly prefer more reliable partners, such monotone partner-quality preferences are typically incompatible with core existence: if coalition values increase with the average reliability of members, then the subset of highest-reliability firms can always profitably deviate from any larger coalition. Banerjee, Konishi, and Sönmez (2001) formalize this mechanism in additively separable hedonic games, showing that when preferences are strictly upward-biased in partner quality, every proposed partition is blocked by a top-ranked subcoalition. In our setting, however, the solvency requirement eliminates the feasibility of such blocking coalitions: small, high-reliability groups cannot satisfy the capital reserve condition needed to cover a potential member default. Thus, regulation removes the very deviations that would otherwise generate an empty core, restoring core existence and selecting a unique stable partition formed by the minimal feasible coalitions of adjacent firms in the reliability ordering.

A central parallel concerns the role of bounded heterogeneity in sustaining stable outcomes. In our environment, insurers differ in risk exposure and capitalization, yet these differences are structured in a way that preserves a consistent ordering of partner attractiveness across coalitions. This prevents reversals in relative desirability when coalition membership changes, ensuring that deviation incentives remain aligned. Gusev, Nesterov, Reshetov, and Suzdaltsev (2024) identify a closely related principle in a non-cooperative allocation setting, where equilibrium existence requires heterogeneity to be sufficiently limited so that no agent gains by unilaterally relocating to a different resource. In both cases, stability is not driven solely by preferences or payoffs, but by the interaction between preference monotonicity and controlled dispersion in agent characteristics. It is this structured, rather than arbitrary, heterogeneity that permits equilibrium or core-stable configurations to exist.

Our work connects to the classical literature on endogenous coalition formation. Demange and Guesnerie (1995) emphasize that coalition feasibility is shaped by institutional or technological constraints, rather than by preferences alone, a perspective that parallels the role of solvency requirements in restricting the set of admissible insurance pools. Ray and Vohra (1999) formalize coalition structures as equilibrium outcomes of deviation-resistant partitioning, and their analysis highlights that stability depends on the set of feasible deviations available to subcoalitions. Our model fits directly into this framework: the reserve requirement endogenously defines feasibility and thereby determines which deviations can undermine or sustain a coalition partition.

Greenberg (1990) develops a general theory of coalition formation in which stability is determined by the set of deviations that are institutionally or structurally admissible. The key insight is that stable coalition structures arise not from assumptions about the agents' foresight or strategic sophistication, but from the rules that govern which coalitional changes are feasible. This perspective is directly relevant to our framework: the solvency requirement specifies the feasible coalition set and thereby determines which deviations can effectively challenge a proposed partition. In this sense, the stable coalition structure in our model is selected by the same mechanism emphasized by Greenberg, namely the restriction of admissible deviations rather than the behavioral assumptions placed on agents.

Hart and Kurz (1983) analyze coalition structures as equilibrium outcomes determined by which blocking coalitions are feasible. This directly parallels our setting, where the solvency requirement filters admissible coalitions and, hence, admissible deviations. The resulting stable partition can be viewed as a coalition-structure equilibrium selected by feasibility and strict reliability ordering.

Bogomolnaia and Moulin (2004) show that when preferences exhibit a dichotomous structure, stability is restored because the set of acceptable coalitions becomes sharply restricted, which prevents profitable deviations by subgroups. In our setting, the solvency requirement induces a similar threshold structure: a coalition is feasible only if its collective reserve can cover the default of a member. This creates an endogenous acceptability set over coalition compositions, effectively reducing coalition comparison to a binary classification between feasible and infeasible groups. As in Bogomolnaia and Moulin's framework, stability arises not from smoothing incentive differences but from the constraint that rules out non-viable coalition deviations. Consequently, the solvency requirement functions analogously to a dichotomous preference boundary, ensuring the existence and uniqueness of the stable coalition partition.

Recent work in network risk-sharing offers a closely related perspective. Kovacevic and Pflug (2022) analyze the formation of risk-sharing structures under feasibility constraints on the allocation of losses and capital, and show that stability is determined by the set of coalitions whose joint exposure remains admissible. Their framework highlights that stable arrangements arise not from unrestricted coalition formation, but from structural constraints that limit profitable deviations. Our solvency requirement plays an analogous role, defining which insurance pools are feasible and thereby selecting the stable coalition partition. Thus, both approaches emphasize feasibility-induced stability in environments with shared risk exposure.

Taken together, these contributions indicate that stable coalition outcomes in risk-sharing environments are determined by the interaction between preference monotonicity and feasibility constraints that delimit admissible deviations. In our setting, this feasibility constraint is given by the solvency requirement, which restricts the set of coalitions whose joint reserve can cover the default of a member. We now formalize this environment. The next section introduces the population of insurers, their risk and capital characteristics, and the regulatory solvency condition that defines the set of feasible coalitions and, consequently, the admissible deviations that determine stability.

3 Model

This section presents a formal model of the reinsurance pooling problem as a hedonic coalition game. Each insurer evaluates potential coalitions based on expected losses, where preferences are entirely determined by the identities and characteristics of coalition partners. We denote the triple $< N, X, p, w, \tau >$ the reinsurance pooling problem.

Where:

$N = \{1, 2, 3, \dots, n\}$ – the set of agents (insurance companies)

p_i – the probability of default of company i, $p = (p_1, p_2, \dots, p_N)$, we assume that bankruptcies are independent.

w_i – the initial capital of company i, $w = (w_1, w_2, \dots, w_N)$

X – the Regulatory Business Intensity (or Capital Intensity Parameter) ($X > 1$)

τ – reserve rate for insurance activities ($\tau < 1$)

Let X denote the Regulatory Business Intensity, the admissible scale of insurance activity per unit of own capital. It captures the regulator-imposed leverage constraint that limits how intensively each firm can utilize its capital in underwriting and reinsurance operations. The value of X is common across firms and reflects the permissible profitability and risk exposure consistent with solvency regulation.

Definition 1: A coalition partition is a set $P = \{S_k\}_{k=1}^{K}$ that partitions N. Thus, $S_k \subset N$ are disjoint and $\bigcup_{k=1}^{K} S_k = N$. The subsets S_k are called coalitions.

Each company can be a member of only one coalition. The utility of each coalition is determined by the formula:

$$u_i(S) = (1 - p_i)Xw_i \left[1 - \sum\nolimits_{L \subseteq S/\{i\}} \min\left(\tau, \frac{\sum_{z \in L} w_z}{\sum_{z \in S/L} w_z}\right) \prod_{z \in L} p_z \prod_{z \in S \setminus (L \cup \{i\})} (1 - p_z)\right] - w_i p_i$$

Here we have L is the set of companies that are in default.

Each insurer seeks to maximize the benefit derived from belonging to a particular coalition S, namely: $\max\limits_{S \subseteq N : i \in S} u_i(S)$.

We consider a decentralized insurance market where firms form mutual pools to stabilize solvency.
Each insurer i holds capital w_i and operates at a common Regulatory Business Intensity X, the admissible scale of insurance activity per unit of capital imposed by solvency regulation. This parameter captures how intensively capital is employed in underwriting and investment.

Definition 1: Feasibility constraint: we assume that S feasible $\Leftrightarrow$ for $\forall\ i \in S$ we have $\tau \sum_{z \in S/\{i\}} w_z \geq w_i$. If the condition does not hold $u_i(S) = 0$.

Assumption 1: We assume that there exists at least one feasible coalition.

The pool must be large enough to fully absorb the defaulting member's capital and the X - scaled share of its insured obligations under a single-default scenario.

When a member defaults, client compensation is funded from two sources: the member's capital base and an additional amount reflecting the business that this capital supported at intensity X. The pool's reserve rate τ does not represent idle cash; it measures the fraction of other members' capital that is exposed through their ongoing operations. Because that capital is also employed at intensity X, the mutual contribution is scaled by X as well. Consequently, the pool finances at most the bankrupt member's capital plus an X -scaled share of partners' committed capacity, implying partial recovery when total liabilities exceed this funding envelope. (Equivalently, for a single default, the payout is bounded by $X(w_i + \tau \sum_{z \in S/\{i\}} w_z)$; full repayment occurs only if this bound exceeds the event liabilities.)

Each insurer i has preferences over all coalitions S that contain i, defined by comparing $u_i(S)$ values. Since preferences depend only on the coalition composition (not external payoffs), this constitutes a hedonic game. We introduce several assumptions that also characterize the game.

Assumption 2: Local homogeneity in probabilities: the Probability of default is generally low ($p \ll \frac{1}{2}$). Companies have approximately the same level of default probability: $p_1 \sim p_2 \sim p_3 \sim \ldots \sim p_N \sim p$ where we have

There exits $0 < \delta \ll 1$ such that $\forall i, j, k \; |p_i - p_j|/p_k < \delta$.

Assumption 3: Local homogeneity in capital level: companies have approximately the same level of capital: $w_1 \sim w_2 \sim w_3 \sim \ldots \sim w_N \sim w$ where we have

There exits $0 < \varepsilon \ll p$ such that $\forall i, j, k \; |w_i - w_j|/w_k < (1 + \varepsilon)(1 + \delta) - 1$.

δ and ε are some small values.

Assumption 4: $\delta < \frac{\varepsilon}{2N}$.

By stability, we will mean the core stability (Bogomolnaia & Jackson (2002).

Definition 2. Coalition partition P is core stable if $\nexists \; T \subset N$ such that for any k: $u_i(T) > u_i(S_k)$ for all $i \in T$.

When a coalition partition P is not core stable, so that $\exists \; T \subset N$ such that for any k: $u_i(T) > u_i(S_k)$ for all $i \in T$ we say that T blocks P.

Assumption 5: Without loss of generality, we assume that $\forall i, j \in N$: $i < j \Longrightarrow p_i w_i < p_j w_j$

Companies prefer partners with lower expected capital losses (with higher reliability).

These assumptions ensure that coalition preferences are consistent and strictly ordered. Without them, changes in coalition composition can reverse preference comparisons in conflicting ways: one insurer may prefer to join a coalition because it contains more reliable partners, while another member of that same coalition may strictly prefer to exclude the entrant for the same reason. Such reversals make coalition preferences incomplete and prevent meaningful stability analysis. The assumptions rule out these inconsistencies and allow us to study the structure of stable pooling arrangements.

Having established the environment and the conditions under which coalition preferences admit a coherent comparison, we now turn to the analysis of coalition formation. The key question is which groupings of insurers, if any, are stable against profitable deviations by their members. We characterize the core-stable coalition structures implied by the model and examine how solvency regulation and heterogeneity in financial reliability shape the resulting pattern of pooling.

4 Results and Discussion

The analysis characterizes the conditions under which stable coalition structures arise in decentralized reinsurance pooling. We show how the solvency constraint and preference monotonicity jointly determine which coalitions can be sustained against deviation, and we identify the unique coalition structure that is stable in the core when these conditions hold.

Lemma 1. *Preferences are anti-monotone in size*: $\forall S \subseteq N, \forall i, j \in S, i \neq j : u_i(S) - u_i(S/\{j\}) < 0$.

In words, once feasibility is satisfied, a smaller coalition is strictly preferred to a larger one by all of its members. The proof of Lemma 1 is provided in Appendix A.

An immediate implication is that any coalition whose reserves are sufficient to cover more than one default is strictly dominated by one of its feasible subcoalitions. Therefore, only minimal coalitions that can cover a single default need to be considered in the stability analysis.

Lemma 2. *Preferences are hedonic and strictly monotone in partner reliability*: when assumption 1-5 are true take any $\delta > 0, \varepsilon > 0$, then $\forall S \subseteq N, \forall i, j \in S, i \neq j$ and $k \notin S$ if $p_k w_k > p_j w_j$ then $u_i(S) > u_i(S/\{j\} \cup \{k\})$.

In words, a coalition is strictly more attractive when it contains members with lower expected default losses. The proof of Lemma 2 is provided in Appendix B.

Together, Lemma 1 and Lemma 2 characterize the structure of preferences over feasible coalitions: insurers strictly prefer (i) coalitions composed of more reliable members and (ii) coalitions that are no larger than required for solvency. These two forces imply that, whenever a feasible coalition contains either a less reliable member or more members than necessary, it is strictly dominated by a smaller feasible coalition composed of more reliable firms. Consequently, any coalition that could serve as part of a stable arrangement must be both minimal in size and contiguous in the reliability ordering. We now show that this structure uniquely determines the core-stable coalition partition.

We formalize this in the following result:

Theorem 1. *Uniqueness of the Core-Stable Partition under Strict Preferences*

Consider the hedonic coalition game induced by the reinsurance pooling problem

$< N, X, p, w, \tau >$.

Each insurer i has capital w_i and default probability p_i.

Assume:

(i) Strict heterogeneity in reliability: $p_1 w_1 < p_2 w_2 < \ldots < p_n w_n$;

(ii) Local homogeneity in probabilities and capital levels (firms forming feasible coalitions are sufficiently similar in p_i and w_i);

(iii) the feasibility constraint (reserve/capacity) is monotone – swapping or removing more reliable members never violates feasibility;

Then there exists a unique core-stable partition of N.

Proof:

Let $r_i := p_i w_i$ with $r_1 < r_2 < \cdots < r_n$. By Assumption (iii), there exists a smallest k such that the coalition $S_1 := \{1, \ldots, k\}$ is feasible.

<u>Step 1</u> (S_1 is the unique top coalition)

Fix $i \in S_1$ and any feasible $S \supseteq \{i\}$. We show that $u_i(S_1) > u_i(S)$, whenever $S \neq S_1$.

Strict improvement by reliability ordering: If S contains any $j > k$, replace j by a more reliable agent $\ell \in \{1, \ldots, k\} \backslash S$ (if any). By Lemma 2 (strict monotonicity in partner reliability), this strictly increases u_i; by Assumption (iii), feasibility is preserved. Iterating this yields a feasible S° with $u_i(S^\circ) > u_i(S)$.

Three cases are possible:

1) $|S^\circ| < k$ is ruled out by minimality of k by Assumption (iii)
2) $|S^\circ| > k$ is impossible by Lemma 1
3) So, we are in the case $|S^\circ| = k$

Minimality: By minimality of k, no proper subset of S_1 is feasible. Therefore, among coalitions S' with $i \in S' \subseteq \{1, \dots, k\}$, feasibility forces $S' = S_1$; in particular, $S^\circ = S_1$. Combining this with strict improvement by reliability ordering yields $u_i(S_1) > u_i(S)$ whenever $S \neq S_1$. Thus S_1 is the unique top coalition, and any core-stable partition P must contain coalition S_1. Otherwise, the members of S_1 would block partition P.

Step 2 (Inductive construction). Remove S_1 and consider the residual game on $N_2 \coloneqq N \setminus S_1$ with the inherited feasibility and preferences. Strict ordering of reliabilities and monotone feasibility are preserved on N_2. Repeating Step 1, there exists a smallest k_2 such that $S_2 \subseteq N_2$ - the minimal feasible coalition formed by the most reliable remaining agents - is uniquely defined and must appear in any core-stable partition of N_2.

Step 3 (Existence and uniqueness). Iterating the argument yields a sequence of disjoint coalitions $S_1, S_2, \dots, S_m$ that covers N (up to infeasible leftovers, if any, which cannot profitably deviate under the same feasibility constraint). Let $P^* \coloneqq \{S_1, S_2, \dots, S_m\}$. By construction, no subset of agents can deviate to a strictly preferred feasible coalition: any such deviation would contradict the defining top-coalition property at some step r. Thus P^* is core-stable. To prove uniqueness, let P^{**} be any core-stable partition. From Step 1, S_1 must be a block of P^{**}. Removing S_1 and applying the same logic to the residual game forces S_2 to be a block, and so on; therefore $P^{**} = P^*$. □

Remark. If strict ordering $r_i = r_j$ holds for some $i \neq j$, the top coalition need not be unique and multiple core-stable partitions may exist; uniqueness requires strict preferences.

The theorem shows that solvency regulation and reliability-based preferences jointly lead to a unique stable configuration of reinsurance pools. Because coalitions must hold just enough capital to cover the default of a single member, any additional member increases exposure without increasing protection, making oversized coalitions unstable. At the same time, insurers strictly prefer partners with lower default risk; thus, any coalition that includes a less reliable member is vulnerable to deviation by a more reliable subset. The only coalitions that can withstand all profitable deviations are therefore the smallest feasible groups formed by consecutive firms in the reliability ranking. This yields a unique stable coalition structure, and no alternative arrangement can be sustained in equilibrium.

5 Conclusion

We studied reinsurance pooling as a hedonic coalition formation problem under a solvency constraint requiring that a coalition be able to absorb the default of one member. Under this constraint, preferences are strictly monotone in partner reliability and strictly anti-monotone in coalition size once feasibility is satisfied. We showed that if insurers are not too heterogeneous in financial reliability, the induced game admits a unique core-stable coalition structure, consisting of minimal feasible groups formed by firms that are adjacent in the reliability ordering. No other coalition configuration can be sustained against deviation.

When heterogeneity in reliability is sufficiently large, coalition preferences cease to be aligned, and no core-stable pooling arrangement exists. Thus, solvency regulation not only restricts individual underwriting capacity but also determines whether decentralized reinsurance markets generate stable risk-sharing structures or require centralized coordination.

Appendix A

Lemma 1. *Preferences are anti-monotone in size*:

$\forall S \subseteq N, \forall i, j \in S, i \neq j$: $u_i(S) - u_i(S/\{j\}) < 0$.

Proof:

This statement $u_i(S) - u_i(S/\{j\}) < 0$ is true if $\frac{u_i(S)-u_i(S/\{j\})}{w_i(1-p_i)X} < 0$. Suppose $|S| = n + 2$.

$$\frac{u_i(S) - u_i(S/\{j\})}{w_i(1-p_i)X} = 1 - \sum\nolimits_{L \subseteq S/\{i\}} \min\left(\tau, \frac{\sum_{z\in L} w_z}{\sum_{z\in S/L} w_z}\right) \prod_{z\in L} p_z \prod_{z\in S(L\cup\{i\})} (1-p_z) -$$

$$-1 + \sum\nolimits_{L \subseteq S/\{i\}} \min\left(\tau, \frac{\sum_{z\in L} w_z}{\sum_{z\in S/L} w_z}\right) \prod_{z\in L} p_z \prod_{z\in S\setminus(L\cup\{i\})} (1-p_z) =$$

Let's assume the coalition is built in such a way that it is possible to cover the loss of capital w_l of only one member, who has probability of default p_l. Then we have:

$$\frac{u_i(S) - u_i(S/\{j\})}{w_i(1-p_i)X} = \tau \prod_{z\in S/\{i\}} (1-p_z) + \sum_1^n p_l \prod_{z\in S/\{l,i\}} (1-p_z)\left(\tau - \frac{w_l}{\sum_{z\in S/\{l\}} w_z}\right) -$$

$$-\tau \prod_{z\in S/\{i,j\}} (1-p_z) - \sum_1^{n-1} p_l \prod_{z\in S/\{l,i,j\}} (1-p_z)\left(\tau - \frac{w_l}{\sum_{z\in S/\{l,j\}} w_z}\right)$$

Because δ and ε are some small values we can consider the simplified expression

$$\tau(1-p)^{n+1} - \tau(1-p)^n + np(1-p)^n\left(\tau - \frac{1}{(n+1)}\right) - p(n-1)(1-p)^{n-1}\left(\tau - \frac{1}{n}\right) =$$

$$= p(1-p)^{n-1}\left((-\tau + n\left(\tau - \frac{1}{(n+1)}\right))(1-p) - (n-1)\left(\tau - \frac{1}{n}\right)\right) =$$

$$= p(1-p)^{n-1}\left(-p\left(-\tau + n\left(\tau - \frac{1}{(n+1)}\right)\right) - \tau + n\left(\tau - \frac{1}{(n+1)}\right) - (n-1)\left(\tau - \frac{1}{n}\right)\right) =$$

$$= p(1-p)^{n-1}\left(p\left(\tau - n\left(\tau - \frac{1}{(n+1)}\right)\right) - \tau + n\tau - \frac{n}{(n+1)} - n\tau + \tau + \frac{n-1}{n}\right) =$$

$$= p(1-p)^{n-1}\left(p\left(\tau - n\left(\tau - \frac{1}{(n+1)}\right)\right) - \frac{1}{n(n+1)}\right)$$

This expression is a linear function of τ therefore, it is enough for us to evaluate the function at the minimum and maximum values of the domain of τ.

1) Because $\tau - \frac{1}{n} \geq 0$, $\tau \in [\frac{1}{n}; \frac{1}{2}]$ let us take $\tau = \frac{1}{n}$. Lust expression would be

$$p(1-p)^{n-1}\left(p\left(\frac{1}{n} - n\left(\frac{1}{n} - \frac{1}{(n+1)}\right)\right) - \frac{1}{n(n+1)}\right) =$$

$$= p(1-p)^{n-1}\left(p\frac{1}{n(n+1)} - \frac{1}{n(n+1)}\right) = p(1-p)^{n-1}(p-1)\frac{1}{n(n+1)} =$$

$$= -p(1-p)^n \frac{1}{n(n+1)} < 0$$

2) let us take $\tau = \frac{1}{2}$. We get

$$p(1-p)^{n-1}\left(p\left(\frac{1}{2}-n\left(\frac{1}{2}-\frac{1}{(n+1)}\right)\right)-\frac{1}{n(n+1)}\right)=$$

$$=p(1-p)^{n-1}\left(p\frac{-n^2+3}{2(n+1)}-\frac{1}{n(n+1)}\right)=$$

$$=p(1-p)^{n-1}\frac{1}{(n+1)}\left(-p\frac{n^2-3}{2}-\frac{1}{n}\right)=$$

$$=p(1-p)^{n-1}\frac{1}{(n+1)}\left(\frac{n^2-3}{2}+\frac{1}{n}\right)(-1)<0$$

Because $n \geq 2$ lust expression is true.

We got the proof of $u_i(S) - u_i(S/\{j\}) < 0$ for the case of one default. Now let S be a coalition whose reserve capacity suffices to cover $m \geq 2$ defaults. Then S is, in particular, feasible for one default. By monotone feasibility, there exists a minimal subcoalition $S^* \subset S$ that is feasible for one default (i.e., any proper subset of S^* violates the one-default constraint). Consider any chain of inclusions

$S^* = S_0 \subset S_1 \subset \cdots \subset S_q = S,$

where each step adds one member and all sets in the chain remain one-default feasible (possible by monotonicity). Apply the result $u_i(S) - u_i(S/\{j\}) < 0$ at each step to obtain, for every incumbent $i \in S^*$, $u_i(S_{t+1}) < u_i(S_t)$ for $t = 0, \dots, q-1$,

hence $u_i(S) < u_i(S^*)$ for all $i \in S^*$. Therefore S^* blocks S relative to any partition containing S, contradicting core stability. It follows that coalitions sized to cover two or more defaults cannot be core-stable, and the analysis may be confined to minimal one-default feasible coalitions. □

Appendix B

Lemma 2. *Preferences are hedonic and strictly monotone in partner reliability*:

when assumption 1-5 are true take any $\delta > 0, \varepsilon > 0$, then $\forall S \subseteq N, \forall i, j \in S, i \neq j$ and $k \notin S$ if $p_k w_k > p_j w_j$ then $u_i(S) > u_i(S/\{j\} \cup \{k\})$.

Proof:

Let us denote $T = S/\{j\} \cup \{k\}$ and $Q = S/\{j\} = T/\{k\}$.

We want to show that $u_i(S) - u_i(T) > 0$. As a result of Lemma 1 we need to consider only the case of default of one company.

$$u_i(S) - u_i(T) =$$

$$= (1-p_i)Xw_i\left[(1-\tau) + \tau\prod_{z\in S/\{i\}}(1-p_z) + \sum_{1}^{|S|-2} p_l \prod_{z\in S/\{l,i\}}(1-p_z)\left(\tau - \frac{w_l}{\sum_{z\in S/\{l\}} w_z}\right)\right] - w_i p_i -$$

$$-(1-p_i)Xw_i\left[(1-\tau) + \tau\prod_{z\in T/\{i\}}(1-p_z) + \sum_{1}^{|S|-2} p_l \prod_{z\in T/\{l,i\}}(1-p_z)\left(\tau - \frac{w_l}{\sum_{z\in T/\{l\}} w_z}\right)\right] - w_i p_i =$$

$$= (1-p_i)Xw_i\left[(1-p_j)\tau\prod_{z\in S/\{i,j\}}(1-p_z) + (1-p_j)\sum_{1}^{|S|-3} p_l \prod_{z\in S/\{i,j\}}(1-p_z)\left(\tau - \frac{w_l}{\sum_{z\in S/\{l\}} w_z}\right)\right] +$$

$$+(1-p_i)Xw_i\left[p_j \prod_{z\in S/\{l,i\}}(1-p_z)\left(\tau - \frac{w_j}{\sum_{z\in S/\{l\}} w_z}\right)\right] -$$

$$-(1-p_i)Xw_i\left[(1-p_k)\tau\prod_{z\in T/\{i,k\}}(1-p_z) + (1-p_k)\sum_{1}^{|S|-3} p_l \prod_{z\in T/\{i,k\}}(1-p_z)\left(\tau - \frac{w_l}{\sum_{z\in T/\{l\}} w_z}\right)\right] -$$

$$-(1-p_i)Xw_i\left[p_k \prod_{z\in T/\{l,i\}}(1-p_z)\left(\tau - \frac{w_k}{\sum_{z\in T/\{l\}} w_z}\right)\right] =$$

$$= (1-p_i)Xw_i[\left(\frac{w_k p_k - w_j p_j}{\sum_{z\in Q} w_z + w_i}\right)\prod_{z\in Q/\{i\}}(1-p_z)$$

$$+ (1-p_j)\sum_{1}^{|S|-3} p_l \prod_{z\in S/\{i,j,l\}}(1-p_z)\left(\tau - \frac{w_l}{\sum_{z\in S/\{l\}} w_z}\right) -$$

$$-(1-p_k)\sum_{1}^{|S|-3} p_l \prod_{z\in T/\{i,k,l\}}(1-p_z)\left(\tau-\frac{w_l}{\sum_{z\in T/\{l\}} w_z}\right)]$$

We will denote the three terms obtained in square brackets, respectively, **A**, **B** and **C**:

$$A=w_i\left(\frac{w_k p_k - w_j p_j}{\sum_{z\in Q} w_z + w_i}\right)\prod_{z\in Q/\{i\}}(1-p_z)$$

$$B=w_i(1-p_j)\sum_{1}^{|S|-3} p_l \prod_{z\in S/\{i,j,l\}}(1-p_z)\left(\tau-\frac{w_l}{\sum_{z\in S/\{l\}} w_z}\right)$$

$$C=w_i(1-p_k)\sum_{1}^{|S|-3} p_l \prod_{z\in T/\{i,k,l\}}(1-p_z)\left(\tau-\frac{w_l}{\sum_{z\in T/\{l\}} w_z}\right)$$

Now we need to consider all possible ratios of capital values and the probability of default for $p_k w_k > p_j w_j$.

Case 1)

Suppose $w_j > w_k$ and $p_k > p_j$. And suppose $w_k(1+\varepsilon)=w_j$, $p_j(1+\delta)=p_k$. Then we have $(1-p_j) > (1-p_k)$ and $\tau-\frac{w_l}{\sum_{z\in S/\{l\}} w_z} > \tau-\frac{w_l}{\sum_{z\in T/\{l\}} w_z}$. So $u_i(S)-u_i(T)>0$

Case 2)

Suppose $w_j < w_k$ and $p_k > p_j$. And suppose $w_k=(1+\varepsilon)w_j$, $p_j(1+\delta)=p_k$.

Because of assumption 5 we can consider two cases.

2.a) Let us consider the situation when the term **A+B** is minimal. It is possible when $\forall l \neq j,k$ we have $p_l=p_k=p_j(1+\delta)$ and $\forall l\neq j,k$ we have $w_l p_l = p_j w_j < p_k w_k$. Then we have $\forall l \neq j,k$ $w_l = w_j/(1+\delta)$.

Let's express through w_j and p_j the expression for $u_i(S)-u_i(T)$.

$$\frac{u_i(S)-u_i(T)}{(1-p_i)X} = w_j\left(\frac{(\varepsilon+\delta+\varepsilon\delta)p_j w_j}{\frac{|S|-1}{1+\delta}w_j}\right)\left(1-p_j(1+\delta)\right)^{|S|-2}+$$

$$+w_j p_j(1+\delta)\left(1-p_j(1+\delta)\right)^{|S|-3}(1-p_j)\left(\tau-\frac{\frac{w_j}{1+\delta}}{\frac{|S|-2}{1+\delta}w_j+w_j}\right)(|S|-3)-$$

$$-w_j p_j(1+\delta)\left(1-p_j(1+\delta)\right)^{|S|-3}\left(1-p_j(1+\delta)\right)\left(\tau-\frac{\frac{w_j}{1+\delta}}{\frac{|S|-2}{1+\delta}w_j+w_j(1+\varepsilon)}\right)(|S|-3)$$

We want to show that this expression >0. This is true if

$$\frac{(\varepsilon+\delta+\varepsilon\delta)(1+\delta)}{|S|-1}\left(1-p_j(1+\delta)\right)+(1+\delta)\left((1-p_j)\tau-\frac{(1-p_j)}{|S|-2+1+\delta}\right)(|S|-3)-$$

$$-(1+\delta)\left(\left(1-p_j(1+\delta)\right)\tau-\frac{\left(1-p_j(1+\delta)\right)}{|S|-2+(1+\varepsilon)(1+\delta)}\right)(|S|-3)>0$$

We know that $\tau \geq \frac{w_j(1+\varepsilon)}{\frac{w_j(|S|-1)}{(1+\delta)}}$ so lust expression must be true when τ is minimal. Let's substitute in the last expression $\tau = \frac{(1+\varepsilon)(1+\delta)}{|S|-1}$.

$$\frac{(\varepsilon+\delta+\varepsilon\delta)(1+\delta)}{|S|-1}\left(1-p_j(1+\delta)\right)+(1+\delta)[\delta p_j\frac{(1+\varepsilon)(1+\delta)}{|S|-1}+\frac{1-p_j(1+\delta)}{|S|-1+\varepsilon+\delta+\varepsilon\delta}-\frac{1-p_j}{|S|-1+\delta}](|S|-3)>0$$

Which is equivalent to the expression

$$\frac{(\varepsilon+\delta+\varepsilon\delta)}{|S|-1}\left(1-p_j(1+\delta)\right)+[\frac{\delta p_j}{|S|-1}+\frac{\delta^2(1+\varepsilon)p_j}{|S|-1}+\frac{1-p_j(1+\delta)}{|S|-1+\varepsilon+\delta+\varepsilon\delta}-\frac{1-p_j}{|S|-1+\delta}](|S|-3)>0$$

Because $\frac{\delta^2(1+\varepsilon)p_j}{|S|-1}$ is relatively small and positive we can ignore it. So lust expression is true if

$$\frac{(\varepsilon+\delta+\varepsilon\delta)}{|S|-1}\left(1-p_j(1+\delta)\right)+\left(1-p_j(1+\delta)\right)[\frac{1}{|S|-1+\varepsilon+\delta+\varepsilon\delta}-\frac{1}{|S|-1+\delta}](|S|-3)>0$$

After simplifications, we get

$$\frac{(\varepsilon+\delta+\varepsilon\delta)}{|S|-1}-[\frac{\varepsilon(1+\delta)}{(|S|-1+\varepsilon+\delta+\varepsilon\delta)(|S|-1+\delta)}](|S|-3)>0$$

Because $\frac{|S|-1}{(|S|-1+\varepsilon+\delta+\varepsilon\delta)(|S|-1+\delta)}(|S|-3)<1$ the lust expression is true if

$\varepsilon+\delta+\varepsilon\delta-\varepsilon(1+\delta)>0$ which is true because $\delta>0$ is true. So $u_i(S)-u_i(T)>0$.

2.b) Let us consider the situation when the term **C** is maximal. It is possible when $\forall l \neq j,k$ we have $p_l = p_j$ and $\forall l \neq j,k$ we have $w_l p_l = p_j w_j < p_k w_k$. From this $\forall l \neq j,k$ $w_l = w_j$.

$$\frac{u_i(S)-u_i(T)}{(1-p_i)X} = w_j\left(\frac{(\varepsilon+\delta+\varepsilon\delta)p_j w_j}{(|S|-1)w_j}\right)\left(1-p_j\right)^{|S|-2}+$$

$$+w_j p_j\left(1-p_j\right)^{|S|-3}\left(1-p_j\right)\left(\tau-\frac{w_j}{(|S|-2)w_j+w_j}\right)(|S|-3)-$$

$$-w_j p_j\left(1-p_j\right)^{|S|-3}\left(1-p_j(1+\delta)\right)\left(\tau-\frac{w_j}{(|S|-2)w_j+w_j(1+\varepsilon)}\right)(|S|-3)$$

We want to show that this expression >0. We know that $\tau \geq \frac{w_j(1+\varepsilon)}{w_j(|S|-1)}$ so lust expression must be >0 when τ is minimal. Let's substitute in the last expression $\tau = \frac{(1+\varepsilon)}{|S|-1}$ and simplify. We get

$$\frac{(\varepsilon+\delta+\varepsilon\delta)}{|S|-1}(1-p_j)+[\frac{\delta(1+\varepsilon)p_j}{|S|-1}+\frac{1-p_j(1+\delta)}{|S|-1+\varepsilon}-\frac{1-p_j}{|S|-1}](|S|-3)>0$$

Lust expression is true if

$$\frac{(\varepsilon+\delta+\varepsilon\delta)}{|S|-1}-[\frac{\varepsilon}{(|S|-1+\varepsilon)(|S|-1)}](|S|-3)>0$$

Because $\frac{|S|-1}{(|S|-1+\varepsilon)(|S|-1)}(|S|-3)<1$ the lust expression is true if

$\varepsilon+\delta+\varepsilon\delta-\varepsilon>0$ which is true because $\delta(1+\varepsilon)>0$ is true. So $u_i(S)-u_i(T)>0$.

Case 3)

Suppose $w_j<w_k$ and $p_k<p_j$. And suppose $w_k=(1+\varepsilon)w_j,\ p_j=p_k(1+\delta)$.

Because of assumption 5 we can consider two cases.

3.a) Let us consider the situation when the term **A+B** is minimal. It is possible when $\forall l\neq j,k$ we have $p_l=p_j$ and $\forall l\neq j,k$ we have $w_lp_l=p_jw_j<p_kw_k$. From this $\forall l\neq j,k\ w_l=w_j$.

Let's express through w_j and p_j the expression for $u_i(S)-u_i(T)$.

$$\frac{u_i(S)-u_i(T)}{(1-p_i)X}=w_j\left(\frac{\left(\frac{1+\varepsilon}{1+\delta}-1\right)p_jw_j}{(|S|-1)w_j}\right)\left(1-p_j\right)^{|S|-2}+$$

$$+w_jp_j\left(1-p_j\right)^{|S|-3}\left(1-p_j\right)\left(\tau-\frac{w_j}{(|S|-1)w_j}\right)(|S|-3)-$$

$$-w_jp_j\left(1-p_j\right)^{|S|-3}\left(1-\frac{p_j}{1+\delta}\right)\left(\tau-\frac{w_j}{(|S|-1)w_j+w_j\varepsilon}\right)(|S|-3)$$

This expression is >0 if the following is true.

$$\frac{\left(\frac{\varepsilon-\delta}{1+\delta}\right)}{(|S|-1)}\left(1-p_j\right)+\left[\tau\left(1-p_j\right)-\tau\left(1-\frac{p_j}{1+\delta}\right)-\frac{1-p_j}{|S|-1}+\frac{1-\frac{p_j}{1+\delta}}{|S|-1+\varepsilon}\right](|S|-3)>0$$

We know that $\tau\geq\frac{w_j(1+\varepsilon)}{w_j(|S|-1)}$ so lust expression must be >0 when τ is minimal. Let's substitute in the last expression $\tau=\frac{(1+\varepsilon)}{|S|-1}$ and simplify. We get

$$\left(\frac{\varepsilon-\delta}{(1+\delta)(|S|-1)}\right)\left(1-p_j\right)+\left[-\frac{\delta p_j(1+\varepsilon)}{(1+\delta)(|S|-1)}-\frac{1-p_j}{(|S|-1)}+\frac{1-\frac{p_j}{1+\delta}}{|S|-1+\varepsilon}\right](|S|-3)>0$$

Let’s simplify this expression step by step:

$$\left(\frac{\varepsilon-\delta}{(1+\delta)(|S|-1)}\right)\left(1-p_j\right)+\left[-\frac{\delta p_j(1+\varepsilon)}{(1+\delta)(|S|-1)}-\frac{1-p_j}{(|S|-1)}+\frac{1-\frac{p_j}{1+\delta}}{|S|-1+\varepsilon}\right](|S|-3)>0$$

$$\left(\frac{\varepsilon-\delta}{(|S|-1)}\right)\left(1-p_j\right)+\left[-\frac{\left(1+\delta-p_j\right)\varepsilon}{(|S|-1+\varepsilon)(|S|-1)}-\frac{\varepsilon\delta p_j}{(|S|-1)}\right](|S|-3)>0$$

$$(\varepsilon-\delta)(1-p_j)-\left[\frac{(1+\delta-p_j)\varepsilon}{(|S|-1+\varepsilon)}+\varepsilon\delta p_j\right](|S|-3)>0$$

$$\varepsilon-\varepsilon p_j-\delta+\delta p_j-(\varepsilon+\varepsilon\delta-\varepsilon p_j)\frac{|S|-3}{|S|-1+\varepsilon}-\varepsilon\delta p_j\frac{|S|-3}{|S|-1+\varepsilon}>0$$

$$\varepsilon\left(1-\frac{|S|-3}{|S|-1+\varepsilon}\right)-\varepsilon p_j\left(1-\frac{|S|-3}{|S|-1+\varepsilon}\right)-\delta+\delta p_j-\varepsilon\delta\frac{|S|-3}{|S|-1+\varepsilon}(1+p_j)>0$$

$$\varepsilon\left(\frac{1+\varepsilon}{|S|-1+\varepsilon}\right)(1-p_j)-\delta+\delta p_j-\varepsilon\delta\frac{|S|-3}{|S|-1+\varepsilon}(1+p_j)>0$$

The lust expression is true if

$$\begin{cases}\varepsilon\left(\dfrac{1+\varepsilon}{|S|-1+\varepsilon}\right)(1-p_j)-\delta>0 & (1)\\ \delta p_j-\varepsilon\delta\dfrac{|S|-3}{|S|-1+\varepsilon}(1+p_j)\geq 0 & (2)\end{cases}$$

From expression (2) we get that $p_j\geq\frac{\varepsilon(|S|-3)}{|S|(1-\varepsilon)-1+3\varepsilon}$

Which is true because of the assumption 2, namely $0<\varepsilon\ll p$.

Expression (1) is true because $\varepsilon\left(\frac{1+\varepsilon}{|S|-1+\varepsilon}\right)(1-p_j)>\frac{\varepsilon}{2|S|}>\frac{\varepsilon}{2N}>\delta$ which is true because of the assumption 3 and 4. So $u_i(S)-u_i(T)>0$

2.b) Let us consider the situation when the term **C** is maximal. It is possible when $\forall l\neq j,k$ we have $p_l=\frac{p_j}{1+\delta}$ and $\forall l\neq j,k$ we have $w_l p_l=p_j w_j<p_k w_k$. From this $\forall l\neq j,k$ $w_l=w_j(1+\delta)$. Let's express through w_j and p_j the expression for $u_i(S)-u_i(T)$.

$$\frac{u_i(S)-u_i(T)}{(1-p_i)X}=w_j\left(\frac{\left(\frac{1+\varepsilon}{1+\delta}-1\right)p_j w_j}{(|S|-1)w_j(1+\delta)}\right)\left(1-\frac{p_j}{1+\delta}\right)^{|S|-2}+$$

$$+w_j\frac{p_j}{1+\delta}\left(1-\frac{p_j}{1+\delta}\right)^{|S|-3}(1-p_j)\left(\tau-\frac{w_j}{(|S|-2)w_j(1+\delta)+w_j}\right)(|S|-3)-$$

$$-w_j\frac{p_j}{1+\delta}\left(1-\frac{p_j}{1+\delta}\right)^{|S|-3}\left(1-\frac{p_j}{1+\delta}\right)\left(\tau-\frac{w_j}{(|S|-2)w_j(1+\delta)+w_j(1+\varepsilon)}\right)(|S|-3)$$

This expression is >0 if the following is true.

$$\frac{\left(\frac{\varepsilon-\delta}{1+\delta}\right)}{(|S|-1)}\left(1-\frac{p_j}{1+\delta}\right)+\left[\tau(1-p_j)-\tau\left(1-\frac{p_j}{1+\delta}\right)-\frac{1-p_j}{(|S|-2)(1+\delta)+1}+\frac{1-\frac{p_j}{1+\delta}}{(|S|-2)(1+\delta)+(1+\varepsilon)}\right](|S|-3)>0$$

We know that $\tau\geq\frac{w_j(1+\varepsilon)}{w_j(|S|-2)(1+\delta)+w_j}$ so lust expression must be >0 when τ is minimal. Let's substitute in the last expression $\tau = \frac{(1+\varepsilon)}{(|S|-2)(1+\delta)+1}$ and simplify. We get

$$\left(\frac{\varepsilon-\delta}{(1+\delta)(|S|-1)}\right)\left(1-\frac{p_j}{1+\delta}\right) + \left[-\frac{\delta p_j(1+\varepsilon)}{(1+\delta)(|S|-2)+1} - \frac{1-p_j}{(|S|-2)(1+\delta)+1} + \frac{1-\frac{p_j}{1+\delta}}{(|S|-2)(1+\delta)+(1+\varepsilon)}\right](|S|-3) = \boldsymbol{Q}$$

We can compare it with expression from 3.a)

$$\boldsymbol{P} = \left(\frac{\varepsilon-\delta}{(1+\delta)(|S|-1)}\right)(1-p_j) + \left[-\frac{\delta p_j(1+\varepsilon)}{(1+\delta)(|S|-1)} - \frac{1-p_j}{(|S|-1)} + \frac{1-\frac{p_j}{1+\delta}}{|S|-1+\varepsilon}\right](|S|-3) > 0$$

And it true that $\boldsymbol{Q} > \boldsymbol{P}$ because $\left(\frac{\varepsilon-\delta}{(1+\delta)(|S|-1)}\right)\left(1-\frac{p_j}{1+\delta}\right) > \left(\frac{\varepsilon-\delta}{(1+\delta)(|S|-1)}\right)(1-p_j)$ and because

$$\left[-\frac{\delta p_j(1+\varepsilon)}{(1+\delta)(|S|-2)+1} - \frac{1-p_j}{(|S|-2)(1+\delta)+1} + \frac{1-\frac{p_j}{1+\delta}}{(|S|-2)(1+\delta)+(1+\varepsilon)}\right] < \left[-\frac{\delta p_j(1+\varepsilon)}{(1+\delta)(|S|-1)} - \frac{1-p_j}{(|S|-1)} + \frac{1-\frac{p_j}{1+\delta}}{|S|-1+\varepsilon}\right]$$

So $u_i(S) - u_i(T) > 0$.□

Bibliography

Abdulkadiroglu Atila and Sonmez T (2003). School choice: A mechanism design approach. "American Economic Review" 93, 729–747.

Abdulkadiroğlu, Atila, Yeon-Koo Che, and Yosuke Yasuda. 2008. “Expanding ‘Choice’ in School Choice.” Economic Research Initiatives at Duke (ERID) Research Paper No. 20.

Alcalde, José, and Antonio Romero-Medina. 2011. “On Stability and Efficiency in School Choice Problems.” Working Paper.

Ashlagi I and Gonczarowski YA (2018). Stable matching mechanisms are not obviously strategy-proof. "Journal of Economic Theory" 177, 405–425.

Bando, Kazuhiro, Ryo Kamei, and Shigeo Muto. 2016. “Two-Sided Matching with Externalities: A Survey.” Journal of the Operations Research Society of Japan 59(1): 35–71.

Banerjee, S., Konishi, H., and Sönmez, T. (2001). Core in a simple coalition formation game. Social Choice and Welfare, 18(1), 135–153.

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

Bogomolnaia, A., and Moulin, H. (2004). Random matching under dichotomous preferences. Econometrica, 72(1), 257–279.

Bourgeois-Gironde, S. (2010). "Regret and the rationality of choices" Philosophical Transactions of the Royal Society B: Biological Sciences, 365, 249–257.

Cerrone, C., Hermstrüwer, Y., and Kesten, O. (2024). School Choice with Consent: An Experiment. The Economic Journal, 134(661), 1760-1805.

Chen Y and Möller M (2024). Efficiency-adjusted deferred acceptance: Consent and behavioral stability in school choice. "Theoretical Economics", forthcoming.

Demange, G., and Guesnerie, R. (1995). Group formation in economics: A survey of the literature. In M. Sertel (Ed.), Advances in Economic Design. Springer.

Dimitrov, D., Lazarova, E., and Sung, S.-C. (2016). Inducing stability in hedonic games. UEA School of Economics Working Paper 2016-09.

Duflo, E., Dupas, P., and Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized experiment in Kenya. American Economic Review, 101(5), 1739-1774.

Dur U and Wiseman T (2019). The stability of school choice with peer effects. "Theoretical Economics" 14, 705–740.

Dur, Umut M., and Tommy Wiseman. 2019. “School Choice with Neighbors.” Journal of Mathematical Economics 83: 101–109.

Dur, Umut M., Timo Mennle, and Sven Seuken. 2018. “First-Choice Maximal and First-Choice Stable School Choice Mechanisms.” Proceedings of the 2018 ACM Conference on Economics and Computation.

Dutta, Bhaskar, and Jordi Massó. 1997. “Stability of Matchings When Individuals Have Preferences over Colleagues.” Journal of Economic Theory 75: 464–475.

Ergin H and Sonmez T (2006). Games of school choice under the Boston mechanism. "Journal of Public Economics" 90, 215–237.

Featherstone C and Niederle M (2016). Boston versus deferred acceptance in an experimental investigation of school choice mechanisms. "American Economic Journal: Microeconomics" 8, 91–117.

Fernandez M (2020). Regret-free truth-telling. "Games and Economic Behavior" 120, 187–202.

Fudenberg D and Levine DK (1998). The theory of learning in games. MIT Press, Cambridge.

Gale D and Shapley LS (1962). College admissions and the stability of marriage. "American Mathematical Monthly" 69, 9–15.

Greenberg, J. (1990). The theory of social situations: An alternative game-theoretic approach. Cambridge University Press.

Gusev, V., Nesterov, A., Reshetov, M., and Suzdaltsev, A. (2024). The existence of a pure-strategy Nash equilibrium in a discrete ponds dilemma. Games and Economic Behavior, 147, 38–51.

Hanushek, E. A., Kain, J. F., Markman, J. M., and Rivkin, S. G. (2003). Does peer ability affect student achievement? Journal of Applied Econometrics, 18(5), 527-544.

Hart S and Mas-Colell A (2000). A simple adaptive procedure leading to correlated equilibrium. "Econometrica" 68, 1127–1150.

Hart, S., and Kurz, M. (1983). Endogenous formation of coalitions. Econometrica, 51(4), 1047–1064.

Kawagoe T and Kojima F (2020). Behavioral stability in matching markets: An experimental study. "Journal of Economic Theory" 188, 105047.

Kesten, Onur. 2010. “School Choice with Consent.” Quarterly Journal of Economics 125(3): 1297–1348.

Kominers, Scott Duke. 2010. “Matching with Preferences over Colleagues Solves Classical Matching.” Games and Economic Behavior 68: 773–780.

Kondratev, Aleksei Y., and Alexander S. Nesterov. 2021. “Minimal Envy and Popular Matchings.” European Journal of Operational Research 292(1): 92–105.

Kovacevic, R., and G. C. Pflug. 2022. “Risk Sharing in Networks: Optimality and Stability.” Mathematics of Operations Research 47 (2): 862–886.

Lazarova, E., Borm, P., and van Velzen, B. (2005). Contracts and insurance group formation by myopic players. CentER Discussion Paper 2005-68.

Lazarova, E., Borm, P., and van Velzen, B. (2009). Contracts and coalition formation based on individual deviations. Queen’s University Management School Working Paper.

Li S (2017). Obviously strategy-proof mechanisms. "American Economic Review" 107, 3257–3287.

Mandler M (2006). Regret and instability in matching markets. "Journal of Economic Theory" 127, 1–20.

Mehta, N. (2018). Time-use and academic peer effects in college. (Working Paper).

Mennle, Timo, and Sven Seuken. 2016. “First Choice-Maximizing School Choice Mechanisms.” Working Paper.

Ortega J and Klein T (2023). Efficiency and manipulability in rank-minimizing school choice mechanisms. "Games and Economic Behavior" 140, 267–286.

Ortega, Javier, and Tobias Klein. 2023. “The Cost of Strategy-Proofness in School Choice.” Games and Economic Behavior 141: 515–528.

Pathak PA and Sonmez T (2008). Leveling the playing field: Sincere and sophisticated players in the Boston mechanism. "American Economic Review" 98, 1636–1652.

Pathak PA and Sonmez T (2013). School admissions reform in Chicago and England: Comparing mechanisms by their vulnerability to manipulation. "American Economic Review" 103, 80–106.

Pycia M and Troyan P (2016). Obvious dominance and random priority. "Journal of Economic Theory" 166, 211–256.

Ray, D., and Vohra, R. (1999). A theory of endogenous coalition structures. Games and Economic Behavior, 26(2), 286–336.

Rees-Jones A (2018). Suboptimal behavior in strategy-proof mechanisms: Evidence from the residency match. "American Economic Review" 108, 224–256.

Revilla, Pablo. 2007. “Many-to-One Matching When Colleagues Matter.” Working Paper.

Sonmez T and Unver MU (2009). Matching, allocation, and exchange of discrete resources. In Arrow K and Intriligator M (eds.) "Handbook of Social Economics". Elsevier, Amsterdam.

Troyan P (2024). Non-obvious manipulability of rank-minimizing mechanisms. "Journal of Mathematical Economics" 113, 103945.

Troyan, Peter, and Thomas Morrill. 2020. “Obvious Manipulations.” Journal of Economic Theory 185.

Troyan, Peter. 2022. “Non-Obvious Manipulability of the Rank-Minimizing Mechanism.” Working Paper.