Lawyer Assignment with Congestion: Redesigning the Comal County Pilot Program

Emre Doğan* Onur Kesten† Danisz Okulicz‡

August 24, 2026

Abstract

The U.S. system for assigning public lawyers to indigent defendants denies choice to manage congestion. A pilot program in Comal County, Texas, granting choice caused overcrowding among popular lawyers. We propose implementing efficient and mobility-free assignments where no defendant prefers to unilaterally switch lawyers. Such assignments always exist and, when feasible, match the unique envy-free allocation. While pilot's mechanism is obviously manipulable, no mobility-free mechanism is, and gains from manipulation are limited. We develop efficient and mobility-free congestion-cutting mechanisms, and offer a practical design preserving choice while guaranteeing defendants prioritizing congestion-level over lawyer identity are never assigned to overburdened lawyers.

JEL classification: C78, D47, D62

Keywords: Lawyer assignment; market design; congestion; mobility-freeness

*Faculty of Economic Sciences, HSE University, Moscow, Russia (edogan@hse.ru)

†School of Economics, The University of Sydney, Sydney, Australia (onur.kesten@sydney.edu.au)

‡ECON Institute, Karlsruhe Institute of Technology, Karlsruhe, Germany (okulicz.danisz@kit.edu).

We would like to thank our Coeditor and three anonymous reviewers for insightful comments that improved the paper. We gratefully acknowledge helpful suggestions and feedback from Anna Bogomolnaia, Alan Miller, Hervé Moulin, William Phan, David Perez-Castrillo, Antonio Romero-Medina, Yu Zhou, Vlad Smirnov, Alex Westkamp, Fuhito Kojima, Utku Unver, Hideo Konishi, and participants at the 2024 SSCW Meeting, 42nd Annual EALE Conference, 24th Annual SAET Conference, XIV LET Conference, MACIE Seminar, 2024 Conference on Mechanism and Institution Design, 9th Hurwicz Workshop, 2026 Australasian Economic Theory Workshop, and 2026 SSCW Meeting.

1 Introduction

In the 75 largest U.S. counties, approximately 80% of criminal defendants are indigent and therefore unable to afford legal representation. Providing legal representation for indigent defendants imposes a significant fiscal burden. In 2012, states alone, excluding the federal government expenditure, spent between $2.2 billion and $2.4 billion annually on indigent defense [Herberman and Kyckelhahn, 2014].[1]

Since the landmark *Gideon v. Wainwright (1963)* Supreme Court ruling in the U.S.,[2] every indigent defendant in a criminal case has a right to be represented by a public attorney. This right, derived from the Sixth Amendment to the U.S. Constitution, is famously encapsulated in the Miranda warning as *"If you cannot afford a lawyer, one will be appointed for you."*

Neither the Sixth Amendment nor the Miranda rights specify *how an indigent defendant's attorney will be appointed* in the United States. By contrast, the right to counsel of choice in criminal proceedings has been internationally recognized since 1966 by the United Nations' *International Covenant on Civil and Political Rights*:

> *"In the determination of any criminal charge against him, everyone shall be entitled to the following minimum guarantees, in full equality, (including) the right to defend himself in person or* ***through legal assistance of his own choosing****."*

Still, the systems for appointing public lawyers vary significantly across countries.[3] For example, while some countries—such as the U.K.—provide defendants with a list of lawyers from which they can choose, others, like Poland and Germany, allow lawyers to decline representation, thereby limiting defendants' options. Within this international spectrum, the American system is characterized by an *extreme denial of choice*.

[1] Current indigent spending figures are estimated to be much larger as the Bureau of Justice Statistics stopped publishing detailed indigent-defense-only totals after 2012. According to the most comprehensive federal figures, states and local governments spent about $66.4 billion on all judicial and legal functions—including prosecution, courts, and public defense—in 2017; public defense is a substantial subset of that total [Buehler, 2021].

[2] The presiding Justice Black stated the rationale of this decision as "*[R]eason and reflection require us to recognize that in our adversary system of criminal justice, any person haled into court, who is too poor to hire a lawyer, cannot be assured a fair trial unless counsel is provided for him... That government hires lawyers to prosecute and defendants who have the money hire lawyers to defend are the strongest indications of the widespread belief that lawyers in criminal cases are necessities, not luxuries.*"

[3] See Aköz et al. [2025] for a detailed overview.

Notwithstanding some regional variations,[4] a typical court in the U.S. retains a list of available lawyers and appoints them to defendants on a *rolling roster* basis; that is, the first case goes to the first lawyer on the roster, the second case to the second lawyer, and so on. Once the number of cases exceeds the number of available lawyers, the procedure repeats, starting from the first lawyer. Defendants are not allowed to express a preference for their lawyers and cannot request the appointment of a specific lawyer, even if that lawyer agrees to represent them.[5]

The main rationale for denying defendants any say in the appointment of their lawyers is the issue of *congestion*. In the words of one of the fathers of Law and Economics, Richard Posner:

> *"The services of the criminal defense bar cannot be auctioned to the highest bidder among the indigent accused—by definition, indigents are not bidders. But these services must be allocated somehow; indigent defendants cannot be allowed to paralyze the system by all flocking to one lawyer."*

Congestion is a major problem in the American legal aid system. In 2007, 73% of county-based public defender offices exceeded the maximum recommended number of cases [Langton and Farole, 2009], with lawyers handling up to 100 cases at any given time [Whitehurst, 2004]. Attorney Jack Talaska, a public defender in Lafayette, LA, gained nationwide attention when it was revealed that on April 27, 2017, he was handling 194 felony cases—an overwhelming workload equivalent to five years' worth of cases.[6][7]

At first glance, the "rolling roster" allocation method appears to provide various advantages in a congested environment: it is simple to implement and administer, and it tends to produce a roughly equal distribution of cases across lawyers. Upon closer examination, however, it has several drawbacks. From a legal perspective, scholars primarily caution against the behavioral effects of denying choice, arguing that it may reduce cooperation between defendants and their lawyers, demotivate lawyers, and

[4]The practice in federal courts and in most state courts is for judges to appoint defense counsel. At the local level, in addition to the rotation system, appointments may also be distributed on an ad hoc basis or delegated to an agency or program, independent of the court system [Nugent-Borakove and Cruz, 2017].

[5]As established by the *Drumgo v. Superior Court* 1973 case.

[6]New York Times, January 2019

[7]This situation is in part due to the scarcity of funds dedicated to legal aid. For example, in 2007, expenditures on public defender offices in the U.S. were lower than expenditures on legal aid in England and Wales alone [Langton and Farole, 2009, 2010, Pyper et al., 2020].

foster perceptions of procedural injustice [Nugent-Borakove and Cruz, 2017, Baker, 2017, Brecher, 1987, Tague, 1974, Schulhofer and Friedman, 1993, Garcia, 1991]. From an economic perspective, this allocation method raises two obvious concerns. First, it may result in an inefficient allocation. For instance, a defendant accused of murder might be assigned to a lawyer experienced only in robbery cases, while a defendant charged with theft could end up with a lawyer specializing in homicide.[8] Similarly, some defendants may be willing to trade lawyer time and attention for higher lawyer quality, while others may be less inclined to accept such compromises. Second, the resulting allocations may be strikingly unfair: experienced and efficient lawyers end up handling roughly the same amount of cases as those who have only recently passed the bar exam. In contrast, a system that assigns more cases to more experienced lawyers could improve the overall quality of legal aid for the majority of indigent defendants. These issues are believed to contribute to a high volume of appeals due to inefficient counsel–the most common reason for appeals in the U.S.– thereby placing an additional burden on an already strained system [Heise et al., 2017].

Dissatisfaction with the existing lawyer assignment system prompted the Texas Indigent Defense Commission and the Justice Management Institute (JMI) to take unprecedented steps in exploring how the right to choose one's own counsel could be extended to indigent defendants in the United States.[9] This initiative aimed to assess its potential impact on defendants, defense attorneys, and the broader criminal justice system. In 2017, this concept materialized through a pilot program called Client Choice, designed and implemented in Comal County, Texas [Nugent-Borakove and Cruz, 2017]. This program granted defendants the flexibility to nominate their preferred lawyers. At the conclusion of the program, JMI determined that Client Choice is a viable model that could be successfully replicated within the typical U.S. criminal justice system.[10]

[8] In addition to specialty, lawyers can often be compared according to certain publicly observable objective quality metrics[Agan et al., 2021, Shem-Tov, 2022, Iyengar, 2007].

[9] A key argument put forth by the proponents of a choice-based model was that it fosters "competition" among lawyers, thereby enhancing the overall quality of legal representation available to indigent defendants.

[10] According to Nugent-Borakove and Cruz [2017], the majority of defendants who were offered the option to choose their attorney did so. Evaluation of the results of the program showed no evidence of adverse effects to defendants, to lawyers, or to the cost effectiveness and administration of the justice system. Client Choice participants pleaded guilty to lesser charges or proceeded to trial more often than their peers. This difference is statistically significant with Client Choice participants 2.96 times more likely than non-Client Choice participants to plead to a lesser charge than to go to trial, even when controlling for prior convictions. Moreover, defendants who chose their lawyers met with their lawyers sooner and more often than their peers, who elected to have the court choose their lawyers through the conventional appointment process.

However, the program encountered several organizational challenges. A key issue was the designers' insufficient attention to managing lawyer caseloads or accounting for defendants' preferences regarding the amount of time they wished to receive from their lawyers. Consequently, as Richard Posner had predicted, the most sought-after lawyers became overwhelmed with an excessive number of cases, while others were left with too few. Nugent-Borakove and Cruz [2017] concluded:

> *"Although the main hypothesis that Client Choice would produce a greater sense of procedural justice is not fully supported, it is possible that it is a function of volume and not Client Choice per se. One unanticipated implementation phenomenon that resulted in several lawyers being inundated with cases early in the launch of Client Choice likely had an impact on the capacity of those lawyers to provide the standard of representation defined in these measures."*

In this paper, we propose managing lawyer congestion using a *direct mechanism design approach*. We model congestion as a discrete resource, and assume that defendants have preferences not only over lawyers who represent them, but rather over pairs of lawyers and their congestion levels, i.e., the number of cases a given lawyer handles at a given moment in time.[11][12] Defendants always prefer less congestion with a given lawyer to avoid competing for lawyer attention. We argue that desirable allocations should satisfy two central properties: (i) *Pareto-efficiency*–no defendant's welfare can be improved without reducing that of another– and (ii) *mobility-freeness*–no defendant can unilaterally improve her welfare by switching to a different lawyer, treating the assignments of others as fixed. The latter property serves as a natural analogue to stability in two-sided matching models, ensuring equilibrium in a non-cooperative sense and fairness in a cooperative sense.

We show that mobility-free and efficient allocations are those that exhibit the least envy among all efficient allocations in the sense that they coincide with the unique envy-free and efficient allocation when one exists. Otherwise, any envy they admit

[11] Unlike Phan et al. [2024] and Tierney [2019], but similarly to Milchtaich [1996], Bogomolnaia and Moulin [2025], and Copland [2023], we consider a discrete model of congestion, that is, each additional case increases the congestion of a lawyer by 1. For more on this point, see Section 6.

[12] Texas law mandates an indigent defendant to be appointed a lawyer in no more than 3 days. While we abstract from dynamic considerations in our modeling, this 3-day mandate indeed makes a static model of lawyer assignment fairly reasonable for practical applications. This motivates the practical implementation of our theory through a batched assignment system we propose in Section 5.2.

is minimal in the sense that it would disappear if the envied defendant's lawyer had just one additional client. While ensuring no envy may not always be feasible under efficiency, an efficient mobility-free allocation always exists.

We then formally assess the lawyer assignment method used in the Comal County pilot program, referred to as the Comal mechanism (CM), through the lens of our formal model. We find that it violates both efficiency and mobility-freeness. It also gives strong incentives for manipulation which are easy for defendants to figure out, i.e., CM is obviously manipulable. The Nash equilibria of the preference revelation game induced by CM need not be mobility-free, when some defendants are not sophisticated.

Motivated by the significant shortcomings of CM, a central objective of our work is to design efficient mechanisms that are mobility-free. This task is challenging because conventional assignment mechanisms cannot be directly applied due to the need to account for agents' congestion preferences. In standard assignment problems, mechanisms typically begin with no initial assignments and progressively allocate resources in rounds based on preferences. However, this approach is unsuitable in our setting since increasing levels of congestion invalidate the preferences of agents assigned a lawyer in earlier rounds, as those preferences are formed under assumptions of lower congestion. To overcome this issue, our mechanisms take the opposite approach: they start from a hypothetical scenario of extreme congestion and systematically reduce congestion levels in a welfare-monotonic manner.

To practically operationalize our proposal, we provide a family of *Congestion-cutting mechanisms* for computing an efficient and mobility-free assignment for any preference profile. A typical member of this family proceeds from a situation in which all lawyers are hypothetically fully congested and, at each step, progressively reduces the congestion levels of underdemanded lawyers. Meanwhile, entitlements are updated at each step to ensure that each client always weakly improves as the algorithm progresses. The process continues until the total hypothetical congestion equals the number of defendants.

Congestion-cutting mechanisms are not only mobility-free but, unlike CM, also immune to obvious manipulations. While no mobility-free mechanism is strategy-proof—a vulnerability reminiscent of the familiar incentive tension for stable mechanisms in two-sided matching—there is a strict limit to gains from manipulation of mobility-free mechanisms. Namely, a defendant who misrepresents her preference can achieve, at best, her most preferred mobility-free assignment. Whereas equilibria of an arbitrary mobility-free mechanism need not be mobility-free, we show that for a subclass of congestion-cutting mechanisms—those that employ individually rational, non-wasteful, and group

strategy-proof stage mechanisms—any pure-strategy Nash equilibrium, whenever one exists, is mobility-free, regardless of the composition of strategic and naive defendants.

Technically, the lawyer assignment problem is a discrete assignment problem with congestible public goods: lawyers are non-excludable but partially rival[13]. Related environments were first studied through the lens of non-cooperative congestion games, where agents choose the resource they want to consume in a decentralized manner. In the setting closest to ours, Milchtaich [1996] shows that a pure-strategy Nash equilibrium always exists; in our setting, this corresponds to a mobility-free assignment. Moreover, a result of Konishi et al. [1997] implies that efficient and mobility-free assignments also exist in our setting.[14] More recent work studies congestion from normative and market-design perspectives, including general assignment problems under congestion preferences [Bogomolnaia and Moulin, 2025], market-design approaches to multiple commons [Tierney, 2019], and school choice under crowding [Phan et al., 2024, Copland, 2023]. Against this background, we study the novel institutional context of assigning public lawyers to indigent defendants and develop direct mechanisms that, in addition to fairness, emphasize efficiency, incentive properties, and practical implementability. We discuss the related literature in more detail in Section 6.

Although our theoretical framework offers a principled solution to assign lawyers while managing congestion, translating it into practice requires addressing two key challenges: the complexity of the preference reporting format and the need to operate within tight legal time constraints, such as the Texas 72-hour deadline for counsel assignment. To overcome these barriers, we propose a simple and scalable implementation plan that combines choice with congestion control, drawing lessons from the Comal County pilot program while improving its design.

Our proposed deployment processes defendants in small, time-bounded batches and presents each with a limited, rotating subset of available lawyers. Defendants express their preferences using intuitive, color-coded bars that represent each lawyer's projected congestion level—green for low, yellow for moderate, and red for high. By ranking these combinations of lawyer and congestion levels, defendants implicitly communicate preferences compatible with our theoretical model. Crucially, the system guarantees *red-zone avoidance*, ensuring that no defendant ends up with a severely overburdened lawyer if

[13]Partially rival in the sense that assigning one defendant to a lawyer does not preclude assigning additional defendants to the same lawyer, but each additional defendant reduces the value of that lawyer to those assigned to him.

[14]We are grateful to an anonymous referee who brought this implication to our attention.

she expresses a clear desire to avoid such outcomes, regardless of others' preferences. This practical implementation strikes a balance between respecting defendants' preferences and maintaining a manageable and equitable distribution of caseloads.

1.1 The Comal County Client Choice Program

The Comal County, Texas pilot program is, to our knowledge, the only attempt to introduce defendants' choice in the American legal aid system. In this subsection we briefly summarize the report on the implementation of the program based on Nugent-Borakove and Cruz [2017]. The program was implemented in Comal County, Texas, during 2016–2017 and involved a total of 54 lawyers handling approximately 3,000 cases per year, with more than 700 cases active at any given time.[15] This caseload suggests that Comal County was relatively uncongested by American standards. The primary motivations for implementing the program included fostering competition among lawyers to boost their motivation, eliminating potential conflicts of interest associated with judicial appointments, and enhancing trust between lawyers and defendants.

Participation in the program was voluntary. Each defendant was given the option to choose to have their counsel appointed in a traditional manner by a judge based on a "rolling roster", or to participate in the Client Choice program where they could state their preferences over lawyers. Defendants who chose to participate in the program were provided with a binder containing lawyer information forms. These forms listed the attorneys available to represent them and included details about each lawyer's educational background and professional experience.[16] As Texas law requires indigent defendants to be appointed counsel within three working days, the program administrators initially intended to give defendants 48 hours to make their decisions; however, jailed defendants were given only 15–20 minutes to review lawyer information and make their choices due to practical constraints. Each defendant was asked to write down the names of 3 lawyers in order of preference. Each defendant was assigned to her *highest ranked available lawyer*.[17] While lawyers were initially allowed to temporarily quit the

[15]Comal County is a mid-sized county of the state of Texas with 129,048 residents at the time of the pilot program.

[16]Specifically, the lawyer information forms included: the lawyer's name; the name and principal address of their law firm; email and website (if available); law school attended and graduation year; year licensed in Texas; types of cases handled (e.g., criminal, domestic relations, etc.); approximate proportion of practice time devoted to indigent criminal defense in the past 12 months; approximate number of criminal defendants represented during that period; any history of public disciplinary actions (with brief explanations, if applicable); and languages spoken in addition to English.

[17]Most defendants were assigned to their first choice lawyers. Despite asking to nominate three

program only due to some unexpected events (e.g., sickness), later the possibility of suspending participation due to excessive workload was added.

The program had some positive effects owing to flexibility provided in choosing counsel. Judges, court administrators, clerks, and prosecutors who were interviewed perceived an improvement in the quality of criminal defense representation.[18] Moreover, defendants rejected harsh plea-deals more often, and received overall lighter sentences. However, defendants reported difficulties in reaching their lawyers, insufficient time devoted to their needs, and overall did not express higher satisfaction with the legal aid provided. The primary driver of these complaints was believed to be lawyer congestion in the resulting assignments. At the onset of the program, over 44% of all cases were assigned to only two lawyers, despite the availability of 52 others. Although this share decreased to approximately 33% over time, the workload was still too high for one of these lawyers, who eventually decided to suspend his participation in the program. While congestion became less severe after six months, a small number of lawyers—between three and five—were consistently handling a third of all cases.[19]

Despite recognizing the success of the pilot program, Nugent-Borakove and Cruz [2017] concluded that, given these concerns, there is a need for improved methods to develop the procedures for implementing a choice model:

> "As a proof of concept, Client Choice in Comal County's assigned counsel system was a success. In addressing the broader question of whether a Client Choice model could be introduced successfully into a U.S. indigent defense system, Comal County's experience suggests that it can. However,

lawyers, the program mainly used the information about the second and third lawyers only if there was some physical reason (sickness, lawyer quitting, etc.) why the defendant could not get their first choice as their counsel.

[18]For example, judges stated that they observed more skilled lawyers to appear in the courtrooms more often. JMI report stated: "[A]nother judge noted that clients seemed to be more satisfied with their lawyers, commenting that there were fewer requests by defendants to fire their lawyers and ask for a substitution of counsel. Another observed that the lawyers known to be the most competent were the ones who were appearing more frequently in their courtrooms. The other lawyers, about whom the judges were worried, were seen far less. One judge spoke to the role Client Choice might play in upholding justice overall: "the delivery of a better product keeps the [justice] system in check," which was supported by another judge's comment that "fewer defendants were as angry with the system as in the past."

[19]According to Nugent-Borakove and Cruz [2017], the reason why a handful of lawyers received a third of all indigent cases at the outset of the program is unclear. More than a third of the lawyers interviewed explicitly commented that defendants based their choices on lawyers' reputations, shaped either by their peers' long-term experiences with those lawyers or by recent high-profile cases the lawyers had won.

as suggested in the Executive Summary of this report, procedures for implementing Client Choice may present different issues and approaches if the defense delivery system involves public defenders and/or contract lawyers instead of assigned counsel...[T]he implementation process in Comal County did raise a few issues that can inform replication of this model in other jurisdictions. For instance, the next jurisdiction that considers this model may want to monitor attorney caseloads to ensure that caseloads do not become unmanageable or overwhelming for an attorney as a result of the number of defendants selecting them."

2 The Model

There is a set of indigent defendants denoted by $N = \{1, 2, ..., n\}$. Each defendant (referred to as female) needs to be represented by a lawyer (referred to as male), and therefore, has to be assigned a lawyer from a set of lawyers $L = \{a, b, c, ...\}$ where $|L| = m$. We use l to represent a generic lawyer. An *assignment* of defendants to lawyers is a mapping $\mu : N \longrightarrow L$. Each lawyer can represent any number of defendants.

A *congestion profile* $s = (s_l)_{l \in L}$ is a vector of non-negative integers such that $\sum_{l \in L} s_l = n$, where s_l denotes the number of clients lawyer l has. Each assignment μ induces a congestion profile s^μ such that for each $l \in L$, $s^\mu_l = |\mu^{-1}(l)|$. When agent i is assigned lawyer l at assignment μ and the total number of clients of l under μ including i is r, we write $\mu_i = (l, r)$. Moreover, we denote the lawyer assigned to i at μ by $\mu(i) = l$.

Defendants are interested not only in the lawyers who represent them, but also in the total number of clients that their lawyer has. Therefore, each defendant has strict preferences $\succ_i$ over ordered $(l, r) \in L \times \{1, ..., n\}$ pairs, where the first component l corresponds to the lawyer and the second component r to the number of l's clients. Let $\succ = (\succ_i)_{i \in N}$ denote a *preference profile*.

Since a lawyer's quality of service can depend on the number of his clients, it is possible that a defendant prefers some lawyer a to another lawyer b for low levels of congestion, while she prefers lawyer b to a for high levels of congestion, e.g., $(a, r) \succ_i (b, r) \succ_i \cdots \succ_i (b, r') \succ_i (a, r')$ where $r < r'$ is a feasible preference for $i \in N$. On the other hand, for each $i \in N$ we rule out $(a, r') \succ_i (a, r)$ where $r < r'$. That is, each defendant prefers sharing her lawyer with a smaller number of other defendants. This

is the only assumption we impose on the preferences of defendants.

Assumption 1. *Defendants prefer lower congestion over higher congestion for any lawyer, i.e., for all $i \in N$, all $l \in L$, and all $r < r'$, we have $(l, r) \succ_i (l, r')$.*

A *lawyer assignment problem* (LAP) is a triple $(N, L, \succ)$ where $\succ_i$ satisfies Assumption 1 for all i. When there is no confusion about the fixed N and L, $\succ$ defines an LAP.

A *mechanism* is a single-valued mapping which yields an assignment for each LAP. Given $\varphi(\succ) = \mu$, we denote i's assignment under profile $\succ$ by $\varphi_i(\succ) = \mu_i$ and i's lawyer at the mechanism by $\varphi_i^l(\succ) = \mu(i)$.

3 Efficiency and fairness

In contrast to the standard assignment context, in the lawyer assignment problem, the objects (here the slots of a lawyer) are unlimited, as lawyers do not have any maximal capacity limits. However, while no defendant can be excluded from the services of any lawyer, each defendant generates a negative externality on others assigned to the same lawyer: as more people get the same object, it becomes less desirable. Hence, lawyers can be interpreted as *congestible public goods*. Since the effects of this externality are already embedded in the preferences of the defendants, the standard notion of efficiency applies.

Definition 1. *An assignment μ' Pareto dominates μ under $\succ$ if no agent is worse off under μ' compared to μ and there is $i \in N$ who strictly prefers her assignment at μ' to that at μ. μ is **efficient** under $\succ$ if it is not Pareto dominated by any feasible μ'.*

At first glance, it may appear reasonable to set a fixed congestion level for each lawyer *a priori* and treat the problem as a standard object allocation with multiple copies (e.g., a school choice problem, where the fixed congestion levels correspond to quotas that must be fully filled at each school). However, this approach fails to ensure efficiency. To illustrate this, consider the following example.

Example 1. ***(A congestion profile with no efficient assignment).*** *$L = \{a, b, c\}$, $n = 6$, and preferences of all agents are the same and as follows:*

$$(a, 1) \succ_i (b, 1) \succ_i (c, 1) \succ_i (a, 2) \succ_i (a, 3) \succ_i (a, 4) \succ_i (b, 2) \succ_i (c, 2)$$

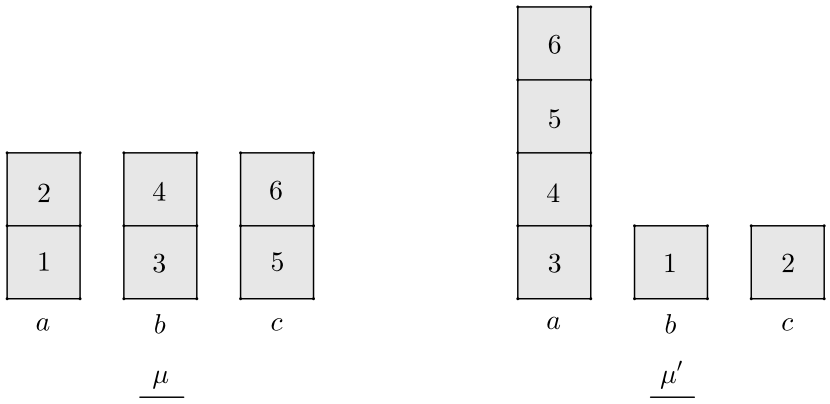


Figure 1: μ and μ' illustrate assignments with congestion profiles $(2, 2, 2)$ and $(4, 1, 1)$, respectively.

Consider a scenario in which lawyer capacities are fixed at $s = (2, 2, 2)$, so that all lawyers have equal congestion levels. The assignment μ, depicted in Figure 1, represents a generic allocation under uniform preferences across defendants. Clearly, all agents are strictly better off under the alternative assignment μ' compared to μ given $\succ$. Since all defendants share the same preferences, for any assignment with the congestion profile $s = (2, 2, 2)$, we can construct an assignment that Pareto dominates it in a similar manner. ♦

In Example 1, μ is not Pareto dominated by any assignment that induces the same congestion profile; however, it is Pareto dominated by μ'. In contrast, it is easy to verify that μ' is efficient. Hence, the congestion profile of the outcome assignment under any efficient mechanism necessarily depends on the preference profile.

For a lawyer assignment system to be considered legitimate, the resulting assignments must not only be efficient but also perceived as fair in practice. Ideally, the assignment should exhibit no envy—every defendant should be content with her assignment when comparing it to those of others.

Definition 2. *A defendant i envies j at the assignment μ under $\succ$ if $\mu_j \succ_i \mu_i$. An assignment is **envy-free** under $\succ$ if no defendant envies another.*

While envy-free assignments always exist, they are often strikingly inefficient. Suppose, for example, that no defendant wishes to be represented by a particular lawyer a, regardless of the congestion profile. If we assign all defendants to lawyer a, any potential envy is trivially eliminated; yet this assignment is unreasonable. Indeed, as shown by Bogomolnaia and Moulin [2025], in our setup, ensuring that an envy-free allocation

is also efficient requires a simple restriction: lawyers who are not assigned any cases should not be considered desirable by defendants given their assignments. That is, an envy-free assignment μ is efficient if and only if $\mu_i \succ_i (l, 1)$ for all $i \in N$ and for all l with $s_l^\mu = 0$. Additionally, Bogomolnaia and Moulin [2025] show that an envy-free and efficient assignment is unique whenever it exists.[20] However, even under very simple and reasonable congestion preferences, an efficient envy-free assignment may, unfortunately, fail to exist.

Example 2. ***(A problem with no efficient and envy-free assignment).*** *$L = \{a, b\}$, $n = 3$, and preferences of all agents are the same and as follows:*

$$(a, 1) \succ_i (a, 2) \succ_i (b, 1) \succ_i (a, 3) \succ_i (b, 2) \succ_i (b, 3)$$

All agents prefer a to b in case they have the same number of clients, and they are ready to accept at most one more congestion level at a compared to b. If all agents are assigned to a, moving one of them to b would make everyone better off. If one or two defendants are assigned to a, each defendant assigned to b envies the one(s) assigned to a. Finally, if all agents are assigned to b, moving all of them to a would be a Pareto improvement. ♦

As efficiency cannot always be reconciled with envy-freeness, we propose an alternative notion of fairness, one that requires no defendant to have an incentive to unilaterally switch to a different lawyer.

Definition 3. *An agent $i \in N$ is mobile at an assignment μ under $\succ$ toward lawyer l if $(l, s_l^\mu + 1) \succ_i \mu_i$. An assignment μ is* ***mobility-free*** *under $\succ$ if no agent is mobile at μ under $\succ$, i.e., for all $i \in N$ and for all $l \in L$, $\mu_i \succ_i (l, s_l^\mu + 1)$.*

The notion of mobility-freeness can be justified in several ways. First, it is based on the idea of equilibrium in a decentralized market where defendants choose their lawyers simultaneously without knowing others' decisions, as in Comal County. As long as all defendants can predict each other's choices perfectly, at an equilibrium, no agent would want to change her lawyer; therefore, each agent prefers her assignment to moving to another lawyer and increasing that lawyer's congestion. In this sense, implementing a

[20] An envy-free assignment satisfying $\mu_i \succ_i (l, 1)$ for all $i \in N$ and for all l with $s_l^\mu = 0$ is called competitive in Bogomolnaia and Moulin [2025]. Moreover, indifferences in the preference of defendants are allowed in their setting. The result we provide here is the projection of Proposition 3 in their work to the case of strict preferences.

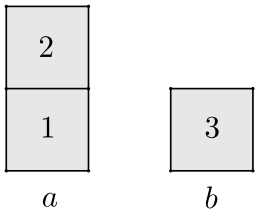


Figure 2: The assignment illustrated above is mobility-free for the problem in Example 2. If a defendant i, currently assigned to lawyer a, decides to move to lawyer b, she would share b with defendant 3. Since $(a, 2) \succ_i (b, 2)$, this would not be a reasonable move. Similarly, defendant 3 is not mobile, since $(b, 1)$ is better than $(a, 3)$, which she would receive if she decided to move to lawyer a. In this example, it is straightforward to verify that an assignment is mobility-free if and only if lawyer a has two clients and lawyer b has one.

mobility-free assignment is procedurally fair. It is equivalent to each defendant exercising her freedom to choose a lawyer while taking the decisions of other defendants as given. Indeed, mobility-free allocations ensure that no farsighted defendant would prefer to switch her lawyer if given the chance.

Second, mobility-freeness can naturally be related to central properties in the standard two-sided assignment and matching problems. Since there is no limit on the number of clients a lawyer can represent in our setting, a mobility-free assignment can be viewed as *non-wasteful*. Moreover, because public defenders are not allowed to express preferences over clients or refuse representation, mobility-freeness coincides with the classical notion of *stability*.

Third, as demonstrated by Bogomolnaia and Moulin [2025], mobility-freeness offers a minimal *ex ante* guarantee to each defendant. Specifically, a mobility-free assignment also satisfies *top-n fairness*: each defendant is guaranteed to do no worse than her nth-ranked lawyer–congestion pair.

Finally, while envy-freeness and mobility-freeness are logically independent, they are related notions.[21] However, under efficiency, every envy-free assignment is mobility-free, and mobility-freeness can be understood as a minimal weakening of envy-freeness. While mobility-free allocations may admit envy, any such envy is borderline in the sense that it would disappear if the envied lawyer had just one additional client. The main result on an abstract congestion game by Milchtaich [1996] implies that mobility-free allocations are guaranteed to exist.

[21] For example, while the assignment where every defendant is assigned to the commonly worst lawyer is envy-free, all defendants are mobile. On the other hand, the assignment illustrated in Figure 2 is mobility-free but not envy-free for the problem in Example 2.

We now turn to the compatibility between efficiency and mobility-freeness.[22] It is straightforward to check that all mobility-free assignments for the problem in Example 2 are also efficient. To see that this need not hold in general, consider an example with two defendants and two lawyers. Defendant 1 prefers being the sole client of lawyer a to any other assignment, and defendant 2 prefers being the sole client of lawyer b to any other assignment. Suppose, moreover, that sharing a lawyer is sufficiently undesirable that each defendant prefers being the sole client of either lawyer to sharing either lawyer with another defendant. Then the only mobility-free assignments are those in which each lawyer serves exactly one defendant. Among these, only the assignment matching 1 with a and 2 with b is efficient. We begin by deriving properties of Pareto improvements over mobility-free assignments.

Lemma 1. *Let μ be a mobility-free assignment under $\succ$. For any $\hat{\mu}$ that Pareto dominates μ under $\succ$, we have*

(i) $s^{\mu} = s^{\hat{\mu}}$, and

(ii) $\hat{\mu}$ is mobility-free.

In words, any Pareto improvement of a mobility-free assignment neither changes its congestion profile nor violates mobility-freeness. Therefore, when verifying the efficiency of a mobility-free assignment, we can treat the problem as one with fixed capacities. That is, it suffices to consider improvements that preserve the same congestion profile.

We omit the straightforward proof of Lemma 1 and provide a brief sketch instead. Let μ be mobility-free under $\succ$ and $\hat{\mu}$ Pareto dominate it. If $\hat{\mu}$ has a different congestion profile, congestion levels for a nonempty set of lawyers, say $L' \subset L$, increased from μ to $\hat{\mu}$. Then, all defendants who are assigned a lawyer in L' under $\hat{\mu}$ are worse off compared to μ whether their lawyers under the two assignments are the same or not, since μ is mobility-free. Therefore, both assignments have the same congestion profile. Furthermore, since no agent is worse off at $\hat{\mu}$ compared to μ, and no agent is mobile at μ, agents would also not be mobile at $\hat{\mu}$ given that the congestion profiles are the same. This shows that $\hat{\mu}$ is mobility-free.

Since Pareto dominance is acyclic and there are only finitely many assignments, Lemma 1 immediately implies parts (i) and (ii) of the following theorem. It also yields

[22]Indeed, this compatibility follows from a result of Konishi et al. [1997]: in a more general noncooperative framework, they prove the existence of strong Nash equilibria, which correspond one-to-one to efficient and mobility-free assignments in our setting. We nevertheless discuss this compatibility explicitly because our results provide clear guidance for finding assignments satisfying both properties in our setting.

a useful characterization. It is straightforward to observe by Lemma 1 that a mobility-free assignment which Pareto dominates every other mobility-free assignment must be efficient. Conversely, suppose μ^* is the unique efficient and mobility-free assignment. Then, all mobility-free assignments must induce the same congestion profile: otherwise, the argument establishing part (ii) would yield efficient and mobility-free assignments with distinct congestion profiles. Therefore, starting from any mobility-free assignment μ, the Pareto-improvement process must terminate at μ^* by Lemma 1. Hence, μ^* Pareto dominates every other mobility-free assignment, establishing part (iii) of the following theorem.

Theorem 1. *For any $\succ$,*

- **(i)** *there exists an efficient and mobility-free assignment,*
- **(ii)** *if an assignment μ is mobility-free, there is an assignment $\bar{\mu}$ with $s^{\bar{\mu}} = s^{\mu}$ such that $\bar{\mu}$ is efficient and mobility-free, and*
- **(iii)** *if an assignment μ^* is mobility-free, then μ^* Pareto dominates every other mobility-free assignment if and only if μ^* is the unique efficient and mobility-free assignment. In this case, $s^{\mu} = s^{\mu^*}$ for every mobility-free assignment μ.*

Theorem 1 has several useful implications. First, if our primary concern is efficiency, then mobility-freeness can be adopted as a viable fairness criterion for assignments. Second, once we know how to find a mobility-free assignment, we can easily derive an efficient assignment without sacrificing mobility-freeness: simply allow the defendants to exchange their lawyers at this assignment. We exploit this fact and provide a family of mechanisms designed to achieve this goal in Section 5.1. Finally, part (iii) identifies the special case in which one mobility-free assignment Pareto dominates all others; this observation will be useful in the incentive analysis below.

Recall that in Example 2, there are multiple mobility-free (and efficient) assignments. However, all such assignments have the same congestion profile. The following example illustrates the fact that more than one congestion profile can support efficient and mobility-free assignments.

Example 3. ***(Multiple congestion profiles supporting efficient and mobility-free assignments).***[23] *$L = \{a, b, c\}$, $n = 5$. Defendant 1 prefers to be represented only by lawyer a, regardless of the level of congestion at other lawyers.*

[23] We use this example to construct a mechanism in Appendix B to show that a defendant may not get her best mobility-free outcome by means of manipulation: a phenomenon we establish in Section 4.

Defendants 2 and 3 prefer to be represented only by lawyers a or b, and for $i \in \{2,3\}$,

$$(b,1) \succ_i (b,2) \succ_i (a,1) \succ_i (a,2) \succ_i (a,3) \succ_i (b,3)...$$

Defendants 4 and 5 prefer to be represented only by lawyers b or c, and for $i \in \{4,5\}$,

$$(b,1) \succ_i (b,2) \succ_i (c,1) \succ_i (c,2) \succ_i (b,3)...$$

There are 6 efficient and mobility-free assignments with 3 different congestion profiles: $s_1 = (3,2,0)$, $s_2 = (2,2,1)$, $s_3 = (1,2,2)$. *In all these assignments, defendant 1 is assigned to lawyer a, two agents from* $\{2,3,4,5\}$ *are assigned to b, those in* $\{2,3\}$ *not assigned to b (if any) are assigned to a, and those in* $\{4,5\}$ *not assigned to b (if any) are assigned to c. The two extreme assignments are as follows:*

$$\mu^{-1}(a) = \{1,2,3\}, \qquad \mu^{-1}(b) = \{4,5\}, \qquad \mu^{-1}(c) = \emptyset$$

$$\hat{\mu}^{-1}(a) = \{1\}, \qquad \hat{\mu}^{-1}(b) = \{2,3\}, \qquad \hat{\mu}^{-1}(c) = \{4,5\}$$ ♦

However, multiplicity arises only if envy cannot be eliminated. Otherwise, there is a unique efficient and mobility-free assignment, which coincides with the necessarily unique efficient and envy-free assignment.[24]

Proposition 1. *Let profile* $\succ$ *admit an efficient and envy-free assignment and* μ *be the (necessarily) unique such assignment. Then* μ *is also the unique mobility-free assignment that is efficient under* $\succ$.

Proposition 1 shows that if the designer aims to guarantee an envy-free and efficient outcome whenever one exists, it suffices to select a mechanism which always yields mobility-free and efficient outcomes.

Definition 4. *A mechanism* φ *is mobility-free (efficient) if, for any preference profile* $\succ$, $\varphi(\succ)$ *is mobility-free (efficient) under* $\succ$.

In fact, in conjunction with Theorem 1, Proposition 1 establishes that mobility-freeness is the strongest fairness notion compatible with efficiency. The following result directly follows from Proposition 1.

Corollary 1. *Any mobility-free and efficient mechanism* φ *yields the necessarily unique envy-free assignment that is efficient whenever it exists.*

[24] To see that the opposite statement is not true, consider the following variant of the problem in Example 2. Let agents 1 and 2 prefer lawyer a to lawyer b regardless of their congestion levels, and let agent 3 have the same preference as in the example. Then, at the unique mobility-free and efficient assignment, agents 1 and 2 are assigned lawyer a and 3 is assigned b, while agent 3 envies 1 and 2. Moreover, it is easy to construct a preference profile allowing for an efficient and envy-free assignment when there are multiple mobility-free (and not necessarily efficient) assignments under the same profile.

4 Comal mechanism vs mobility-free mechanisms

In this section, we first revisit the Comal County pilot program and ask how the Comal mechanism (CM) performs through the lens of our formal framework. After discussing important shortcomings of CM in terms of mobility-freeness and incentives, we pose the same questions for mobility-free mechanisms.

While offering the flexibility to indicate preferences over lawyers, the administrators of the Comal County pilot program did not anticipate the scale of arising congestion in lawyer assignments. Consequently, they did not elicit the preferences of defendants under differing levels of workload of lawyers. Instead, defendants were simply asked to provide an ordered list of their top three lawyers irrespective of other characteristics of the market in which they participate. Each defendant was then assigned to her highest-ranked lawyer so long as the lawyer was available. An important point worth emphasizing is that the assignment mechanism in Comal County is not a direct mechanism in the formal sense that a mechanism is defined in our model. This point requires additional assumptions to be recast in our setting.

When described in the above fashion, defendant preference reports in Comal County are incomplete in the sense that defendants are asked to rank-order lawyers under the presumption that the nominated lawyers are going to be uncongested, i.e., any lawyer will be available to represent only the client that chooses him.

Toward a formal analysis, one way to recreate the Comal County assignment system within our framework is to assume that defendants report only those portions of their preferences when each lawyer's congestion level is set to 1. This enables us to reformulate the Comal mechanism in our setting as follows.

The Comal mechanism (CM): *Each defendant $i \in N$ has a preference $\succ_i$ satisfying Assumption 1 and defendants report only preferences over $(l, 1)$ pairs for all $l \in L$. Assign each defendant her favorite lawyer.*

We begin by underlining two important issues with CM. First, it is easy to see that CM is not mobility-free, e.g., a client who realizes that her favorite lawyer—assuming she were his only client—is in fact highly congested would rather switch to a less congested lawyer. Second, such a switch is also fairly straightforward to achieve strategically, which consequently makes CM *obviously manipulable*. We illustrate this in the example following the formal definition of this deficiency.

Definition 5. *Fix a mechanism φ. For any defendant i, any true preference $\succ_i$, and*

any reported preference $\hat{\succ}_i$, let $\bar{\mu}_i^{\varphi}(\hat{\succ}_i;\succ_i)$ denote the ***best-case scenario*** *outcome for defendant i when she reports $\hat{\succ}_i$, evaluated according to her true preference $\succ_i$, if there exists some $\succ_{-i}$ such that:*

- $\bar{\mu}_i^{\varphi}(\hat{\succ}_i;\succ_i) = \varphi_i(\hat{\succ}_i,\succ_{-i})$, *and*
- *there is no $\succ'_{-i}$ such that $\varphi_i(\hat{\succ}_i,\succ'_{-i}) \succ_i \bar{\mu}_i^{\varphi}(\hat{\succ}_i;\succ_i)$.*

Similarly, let $\underline{\mu}_i^{\varphi}(\hat{\succ}_i;\succ_i)$ denote the ***worst-case scenario*** *outcome for defendant i when she reports $\hat{\succ}_i$, evaluated according to her true preference $\succ_i$, if there exists some $\succ_{-i}$ such that:*

- $\underline{\mu}_i^{\varphi}(\hat{\succ}_i;\succ_i) = \varphi_i(\hat{\succ}_i,\succ_{-i})$, *and*
- *there is no $\succ'_{-i}$ such that $\underline{\mu}_i^{\varphi}(\hat{\succ}_i;\succ_i) \succ_i \varphi_i(\hat{\succ}_i,\succ'_{-i})$.*

A mechanism φ is ***obviously manipulable*** *if there exists a defendant i, a true preference $\succ_i$, and a reported preference $\hat{\succ}_i$ such that i improves her best-case scenario outcome or worst-case scenario outcome by reporting $\hat{\succ}_i$ instead of $\succ_i$, i.e., $\bar{\mu}_i^{\varphi}(\hat{\succ}_i;\succ_i) \succ_i \bar{\mu}_i^{\varphi}(\succ_i;\succ_i)$ or $\underline{\mu}_i^{\varphi}(\hat{\succ}_i;\succ_i) \succ_i \underline{\mu}_i^{\varphi}(\succ_i;\succ_i)$.*

Example 4. *Let $n = 3$ and $L = \{a, b, c\}$. Consider the following preference profile: for all $i \in \{1, 2, 3\}$, $\succ_i$: $(a, 1) \succ_i (b, 1) \succ_i (b, 2) \succ_i (b, 3) \succ_i (c, 1) \succ_i (a, 2) \succ_i (c, 2) \succ_i (c, 3) \succ_i (a, 3)$.*

CM effectively elicits only the restricted ordering $(a, 1) \succ_i (b, 1) \succ_i (c, 1)$, as a result of which each defendant is assigned to lawyer a. This assignment clearly violates mobility-freeness, since each defendant would rather be served alone by lawyer b.

Take any defendant. If she reports truthfully (declaring lawyer a as her first choice), the worst-case scenario happens when everyone else also reports truthfully. In this case, all defendants get $(a, 3)$. If she instead reports b as her first choice, the worst-case scenario happens when everyone else reports the same way. However, in that case, the defendant gets the much better assignment $(b, 3)$. ♦

Given the strong incentives to game the Comal mechanism, it is natural to explore the Nash equilibrium outcomes of the induced preference revelation game and to ask whether our fairness notion can be achieved under equilibrium play. At the same time, since an equilibrium analysis nonetheless relies on the assumption that defendants are strategically sophisticated, it is worth questioning the validity of such an assumption in the present context.

In fact, Nugent-Borakove and Cruz [2017] suggest that when declaring preferences, defendants were not fully contemplating other defendants' strategies:

> "Based on the interviews with attorneys and judges, Client Choice defendants tended to select "popular" lawyers, which reportedly resulted in these lawyers being busier and perhaps less likely to respond promptly to phone calls. This early tendency to gravitate toward a small number of lawyers could explain why defendants may not have felt their lawyer met with them when asked."

We infer that the defendants failed to predict that if a given lawyer is well liked by them, then he will be well liked by other defendants and will be very congested. This suggests that not all defendants may be strategically sophisticated.[25]

Below we analyze the (complete information) Nash equilibria of the preference revelation game induced by the Comal mechanism.[26] In such a game, a defendant's strategy is a preference report possibly different from her true preference ordering while the outcomes are evaluated according to true preferences. Motivated by the above discussion, we argue that defendants are likely to be heterogeneous in sophistication, with differences in education, experience, and familiarity with legal processes leading some to report preferences truthfully while others may behave strategically. Unequal access to legal advice and informal guidance further reinforces this divide, as better-informed defendants are more likely to understand and exploit gaming options. At the same time, cognitive constraints, time pressure, and the stress of legal proceedings make truthful reporting a realistic benchmark for many participants. This distinction also reflects institutional reality, where first-time defendants are more likely to be naive, while repeat participants are more likely to act strategically.

To formalize this distinction, and although we first apply it to the Comal mechanism, it is convenient to define the preference revelation game for a direct mechanism φ in general. Following Pathak and Sönmez [2008], we distinguish between two types of

[25] In Comal County, defendants were not informed of lawyers' current congestion levels. Although this might seem like a reasonable design improvement, due to the dynamic nature of the process, choosing less congested lawyers requires strategic foresight, as many newcomers may later select the same lawyer. Moreover, less strategic defendants might misinterpret this information as a guaranteed low congestion level, which in turn undermines perceptions of procedural fairness. In Section 5.2, we propose an alternative practical design enabling dynamic lawyer assignment without burdening defendants with the information of current congestion while maintaining promised congestion levels.

[26] Throughout this work, we restrict attention to pure-strategy Nash equilibria.

defendants. A **naive** defendant reports her true preferences.[27] Hence, her strategy set is a singleton. Let $K \subseteq N$ be the set of naive defendants. A **sophisticated** defendant best-responds to others. Therefore, under a direct mechanism, the strategy set of such a defendant consists of all strict orderings of lawyer–congestion pairs.[28] All of the above—including the identity of naive defendants and the fact that their strategy sets are singletons—is common knowledge. We denote the preference revelation game induced by an underlying true preference profile $\succ$, a set of naive agents $K \subseteq N$, and a direct mechanism φ by $\Gamma^{\varphi}(K, \succ)$.

Proposition 2. *A Nash equilibrium outcome of the preference revelation game induced by the Comal mechanism need not be a mobility-free assignment when there is at least one naive defendant, i.e., for any set of naive defendants $K \neq \emptyset$, there is a preference profile $\succ$ such that the game $\Gamma^{CM}(K, \succ)$ has a Nash equilibrium outcome assignment that is not mobility-free under $\succ$.*

Proposition 2 brings further bad news for the Comal mechanism. While a mobility-free assignment would be a desirable scenario under the Comal mechanism, Proposition 2 suggests that this is a tall order: such a situation is achievable only when all defendants can jointly coordinate.[29]

We now turn our attention to mobility-free mechanisms. Recall that mobility-freeness corresponds to stability in our context. Although ours is a one-sided assignment setting, unlike the two-sided matching where agents on both sides have preferences, we show below that mobility-free mechanisms and stable mechanisms in classical two-sided models share several important strategic commonalities.

A curious question at this point is whether the incentive shortcoming of CM, discussed earlier, can be overcome by switching to a mobility-free mechanism. The next result gives an affirmative answer.

[27] Haltiwanger and Waldman [1985] study the equilibrium of a stylized model in which agents choose between two paths that exhibit congestion effects. Although they use similar terminology to distinguish between agent types, their notion of sophistication is informational rather than strategic: naive agents are limited in their ability to form expectations and thus hold incorrect beliefs, whereas strategic agents possess rational expectations.

[28] For the Comal mechanism, this reduces to strict orderings of lawyers at unit congestion. Since CM uses only the defendant's top reported lawyer, one can equivalently think of the strategy of a sophisticated defendant as the choice of a top lawyer.

[29] To illustrate that a Nash equilibrium of the Comal mechanism can violate mobility-freeness when some defendants are allowed to be naive, reconsider Example 4 and let only agent 1 be strategic, that is $K = \{2, 3\}$. In the unique Nash equilibrium, defendants 2 and 3 are assigned lawyer a and 1 is assigned b. This assignment is not mobility-free since defendants 2 and 3 are mobile toward lawyer c.

Proposition 3. *No mobility-free mechanism is obviously manipulable.*[30]

Troyan and Morrill [2020] showed that stable mechanisms are robust to obvious manipulations in two-sided matching. We established above that mobility-free mechanisms possess the same property: no defendant can ensure a better best- or worst-case scenario outcome by misreporting.

We next investigate whether it is possible to achieve an even stronger incentive property than immunity to obvious manipulation. The natural candidate is the well-known notion of strategy-proofness. It turns out that strategy-proofness is unfortunately incompatible with mobility-freeness, a phenomenon reminiscent of the incompatibility between stability and strategy-proofness in two-sided matching.

Definition 6. *A mechanism φ is strategy-proof if no defendant can benefit by unilaterally misrepresenting her preference, i.e., for all $\succ$, for all $i \in N$, and for all $\succ_i'$, either $\varphi_i(\succ) \succ_i \varphi_i(\succ_i', \succ_{-i})$ or $\varphi_i(\succ) = \varphi_i(\succ_i', \succ_{-i})$.*

Theorem 2.

(i) *No mobility-free mechanism is strategy-proof.*[31]

(ii) *For any mobility-free mechanism φ, a manipulating defendant can get no better lawyer assignment than at her best mobility-free assignment, i.e., for all $i \in N$, $\succ$, and $\succ_i'$, there exists an assignment μ that is mobility-free at $\succ$ such that either $\mu_i = \varphi_i(\succ_i', \succ_{-i})$ or $\mu_i \succ_i \varphi_i(\succ_i', \succ_{-i})$.*

Theorem 2 also clarifies that the incompatibility between mobility-freeness and strategy-proofness does not rely on the existence of multiple mobility-free congestion profiles. As the proof in the Appendix shows, the incompatibility already arises in the problem of Example 2, where all mobility-free assignments induce the same congestion profile. By contrast, Theorem 1(iii) together with Theorem 2(ii) implies a useful sufficient condition for nonmanipulability at a profile: if there is a unique efficient and mobility-free assignment under $\succ$, even if there are multiple mobility-free assignments, no efficient and mobility-free mechanism is manipulable at that profile. Indeed, such

[30]The lack of obvious manipulations is related to the top-n guarantee property of mobility-free allocations proposed by Bogomolnaia and Moulin [2025], which states that at any mobility-free allocation, any defendant must obtain at worst one of their top-n-ranked alternatives. In fact, the relationship is even stronger. For any fixed preference of a given defendant, and for any top-n outcome for this defendant, there exists a preference profile for which this outcome is achieved by the defendant at *any* mobility-free assignment.

[31]Indeed, Example 2 we use in the proof of this result also reveals that no top-n fair mechanism is strategy-proof.

a mechanism must select the unique efficient and mobility-free assignment at the true profile. Since this assignment Pareto dominates every other mobility-free assignment by Theorem 1(iii), any outcome attainable by a unilateral misreport is bounded above by this assignment by Theorem 2(ii). More generally, part (ii) identifies a defendant's most preferred mobility-free assignment as an upper bound on what she can obtain by manipulation. This upper bound need not be tight: for some mobility-free mechanisms and profiles, some defendants cannot obtain their most preferred mobility-free assignment by any manipulation.[32]

We defer our equilibrium analysis of mobility-free mechanisms to Section 5.1. We note, however, that pure-strategy Nash equilibria of an arbitrary mobility-free mechanism need not be mobility-free even if the mechanism is efficient.[33] Motivated by this observation, we introduce a set of mobility-free mechanisms whose equilibria are *always* mobility-free. When some agents are naive, existence of a pure-strategy Nash equilibrium is no longer guaranteed.[34] However, unlike the Comal mechanism, the mechanisms we propose guarantee that every equilibrium, whenever one exists, is mobility-free regardless of the composition of naive and sophisticated defendants.

5 Redesigning the Comal County Program

Our main goal in this paper is to propose a market design approach for reforming the lawyer assignment program, aiming to provide a blueprint that can be adapted by other cities across the United States. First, in our theoretical analysis, we progressively develop a class of mobility-free and efficient mechanisms with good incentive properties subject to our desiderata. Second, we develop a practical framework to address implementation challenges, particularly those arising from the language of preference reporting and the dynamic nature of real-world deployment. We consider each aspect of the problem in the next two subsections.

[32]This contrasts with the standard two-sided matching benchmark, where the corresponding bound is tight: an agent who does not receive her best stable outcome can obtain it by truncating her preference list below that outcome. Example 6 in Appendix B shows that this tightness property need not hold in our setting.

[33]See Example 7 in Appendix B.

[34]See Example 8 in Appendix B.

5.1 Congestion-cutting algorithms: Theory

Despite studying the compatibility between mobility-freeness and various other desiderata, we are yet to provide a mobility-free mechanism. Since defendants confer an externality on each other in our context, a straightforward 'over-the-counter' approach that directly applies existing assignment mechanisms proves to be elusive.[35] The primary reason is that, to our knowledge, all widely studied assignment mechanisms and their variants tend to gradually "increase congestion". As these algorithms proceed and more agents are assigned, congestion intensifies, potentially making earlier assignees worse off. Of course, this is irrelevant for existing assignment applications as congestion does not matter. A major goal in our design is to deal with this subtlety by developing mechanisms that gradually reduce congestion, thereby allowing us to "keep track of" assignees' welfare across rounds as the process unfolds.

Example 1 demonstrates that a lawyer assignment problem cannot be immediately reduced to a fixed-capacity problem. On the contrary, Theorem 1 suggests that the ability to efficiently assign defendants at a fixed-capacity problem remains useful provided that the "right" capacities can be found. In this section we reconcile the two perspectives and propose a way of adjusting *fixed-capacity problems*. To that end, we start by formally defining a fixed-capacity problem, a solution to this problem, and a mechanism.

Definition 7. *A fixed-capacity problem (FCP) is a slot reassignment problem without congestion preferences defined as follows: Fix N and L. Each $i \in N$ is endowed with exactly one slot at some lawyer. Let $\omega_i \in L$ denote i's endowment and ω be an endowment vector. Each $l \in L$ has a fixed number of slots (capacity) q_l which is not less than the total number of defendants who are endowed with a slot at l, i.e., $|\{i \in N : \omega_i = l\}| \leq q_l$. $(q_l)_{l \in L}$ denotes a capacity vector. Each $i \in N$ has a strict preference P_i over the set of lawyers L, and $P = (P_i)_{i \in N}$ is a preference profile. Therefore, fixing N and L, an FCP is a triple $[\omega, q, P]$.*

Note that as we fix N and L, an FCP is equivalent to an object reallocation problem (or, allocation problem with endowments), where objects (lawyers) potentially have multiple copies (slots), each agent is endowed with exactly one copy of some object,

[35] Our point here concerns direct mechanisms in the matching and assignment literature. It is worth noting, however, that the game-theoretic literature provides simple, polynomial-time procedures for dynamically reaching a pure-strategy Nash equilibrium in congestion games of the kind considered here. In particular, the constructive proof of existence in Milchtaich [1996] induces such a procedure.

and agents have preferences over the objects. Then, by construction, only a capacity vector satisfying $n \leq \sum_{l \in L} q_l$ could be associated with an FCP, that is, the total number of objects to be allocated is not less than the number of agents.

Definition 8. *A solution to a fixed-capacity problem is a reassignment of defendants to lawyers, where no lawyer l is assigned more than q_l defendants. A mechanism ψ is a mapping that yields a solution to each problem. $\psi_i[\omega, q, P] \in L$ is i's lawyer assignment for the problem $[\omega, q, P]$.*

We now endeavor to define a family of algorithms that generate mobility-free and efficient assignments. Any member is essentially characterized by the choice of the mechanism ψ that solves an FCP at each round and is called the *congestion-cutting ψ mechanism*, or simply CC^{ψ}.[36] Independent of the choice of mechanism ψ, the algorithm adjusts the FCP that ψ solves at each step by modifying the endowments of the defendants, the capacity of lawyers, and preferences of defendants (P_i) based on their original congestion preferences ($\succ_i$).

To fix ideas, we first provide an informal description of a typical algorithm. In the first step, the capacity of each lawyer is set at the highest possible level n, and the algorithm considers the portion of defendants' original preferences $\succ$ associated with the congestion profile of $q^1 = (n, n, ..., n)$. This amounts to assuming that each defendant i compares the lawyers in P_i^1 according to her preferences in $\succ_i$ as if each lawyer l were ultimately assigned n defendants and thus operated at full capacity. Since there is no capacity constraint on the solution to the problem in the first step—where all agents could be assigned to the same lawyer—we arbitrarily fix the initial endowment vector to $\omega^1 = (l_1, l_1, ..., l_1)$, which is of no consequence for the final outcome.

In general, for any capacity profile q, $\succ_i$ induces a strict preference ordering over lawyers, which we denote by $\succ_i^q$; that is, $a \succ_i^q b$ if and only if $(a, q_a) \succ_i (b, q_b)$ for any pair of lawyers a and b. The solution to the problem in Step 1 is simply $\psi[\omega^1, q^1, P^1] = \psi[\omega^1, q^1, \succ^{q^1}]$. At the beginning of any other step k, the algorithm first considers the underdemanded lawyers that are not at full capacity with respect to q^{k-1} in the previous step, reduces their capacity by one, and leaves the capacity of the remaining lawyers unchanged. Once the capacity at the beginning of step $k \neq 1$ is set, say to some q^k, the preference of each defendant i at this step (P_i^k) is set to $\succ_i^{q^k}$. Finally, the assigned

[36] In principle, a different fixed-capacity mechanism can be used at each step of the algorithm as long as each satisfies the properties that guarantee an efficient mobility-free final outcome.

lawyers at each step become the endowments for the corresponding defendants at the next step, i.e., for each $k \neq 1$, $i \in N$, $\omega_i^k = \psi_i[\omega^{k-1}, q^{k-1}, P^{k-1}]$.

Algorithm 1 (**The congestion-cutting ψ mechanism (CC^ψ)**). *Fix $(N, L, \succ)$.*

Step 1. *Set $q^1 = (n, n, ..., n)$. For all $i \in N$, set $P_i^1 = \succ_i^{q^1}$ and $\omega_i^1 = l_1$. The tentative assignment for the stage problem is $\psi[\omega^1, q^1, P^1]$.*

Step $k > 1$***.*** *Set $q_l^k = q_l^{k-1}$ if the capacity of l is full in step $k-1$, otherwise $q_l^k = q_l^{k-1} - 1$. For all $i \in N$, set $P_i^k = \succ_i^{q^k}$ and $\omega_i^k = \psi_i[\omega^{k-1}, q^{k-1}, P^{k-1}]$. The tentative assignment for the stage problem is $\psi[\omega^k, q^k, P^k]$.*

The algorithm terminates at the step k^* where the total capacity is equal to $\sum_{l \in L} q_l^{k^*} = n$. The final outcome is $CC^\psi(\succ) = \psi[\omega^{k^*}, q^{k^*}, P^{k^*}]$. ♦

The algorithm is well-defined since at each step it solves an FCP and, by construction, we reach a step such that the total capacity is equal to n. Therefore, the final outcome is a solution to a lawyer assignment problem. Moreover, as long as ψ solves fixed-capacity problems in polynomial time, the algorithm runs in polynomial time as the total capacity starts at $m \times n$ and strictly decreases at each consecutive step.

Toward reaching optimal assignments via a congestion-cutting algorithm, we first define two key properties of fixed-capacity mechanisms that are standard in the literature. The first condition requires that a lawyer should serve any client who prefers him, unless he is at full capacity.

Definition 9. *A fixed-capacity mechanism ψ is non-wasteful if no defendant prefers an underdemanded lawyer with a free slot to the lawyer she is assigned, i.e., if l is such that $|\psi_l^{-1}[\omega, q, P]| < q_l$, there exists no i such that $l \; P_i \; \psi_i[\omega, q, P]$.*

The second property, essential for ensuring efficiency in the final outcome, is the efficiency of the fixed-capacity mechanism used in each step.

Definition 10. *A fixed-capacity mechanism ψ is efficient if for any problem $[\omega, q, P]$, $\psi(\omega, q, P)$ is not Pareto dominated by any assignment that is a solution to $[\omega, q, P]$.*

Using the simplest efficient (and thus non-wasteful)[37] fixed-capacity mechanism from the literature, we next illustrate the operation of a simple congestion-cutting algorithm

[37]Efficiency implies non-wastefulness in a fixed-capacity problem without congestion preferences: if a defendant prefers an underdemanded lawyer l with a free slot to her assigned lawyer, moving her to l hurts no one.

based on a *serial dictatorship*. This example will also help us identify an additional requirement to ensure a mobility-free outcome.

Example 5. *Let* $n = 3$ *and* $L = \{a, b, c\}$. *Consider the following preference profile: for all* $i \in \{1, 2\}, \succ_i$: $(a,1) \succ_i (a,2) \succ_i (b,1) \succ_i (b,2) \succ_i (a,3) \succ_i (b,3) \succ_i \cdots$ *and* $\succ_3$: $(a,1) \succ_3 (a,2) \succ_3 (a,3) \succ_3 (b,1) \succ_3 (b,2) \succ_3 (b,3) \succ_3 \cdots$. ψ *is the serial dictatorship mechanism defined for the order of agents* 1, 2, *and* 3, *i.e., given lawyers' capacities, defendant* 1 *selects her most preferred lawyer among those with available slots. Then, defendant* 2 *chooses her best available lawyer given* 1*'s choice. Finally, defendant* 3 *chooses her best lawyer with a free slot given the choices made by* 1 *and* 2.

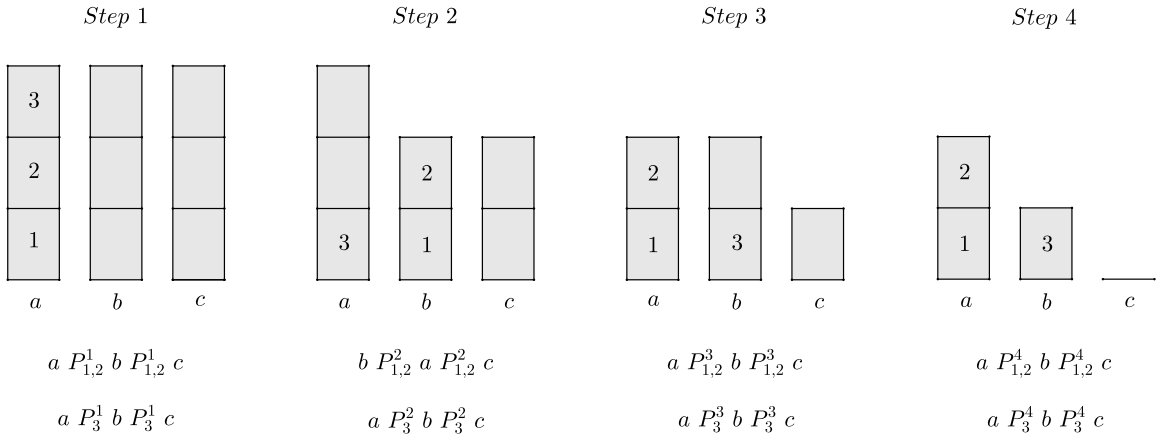


Figure 3: The number of boxes shown for each lawyer indicates his capacity at the given step. P_i^k is the adjusted preference of agent i that is derived from $\succ_i$ based on the lawyers' capacities at step k. The placement of agents represents the assignment given by ψ at the end of each step.

Observe that defendant 3 *is lawyer–a–lexicographic, meaning that she strictly prefers lawyer a over any other lawyer at every step of the algorithm. Moreover, no defendant is ever willing to be assigned to lawyer c, regardless of the lawyers' capacities. In step 1, each lawyer has a capacity of three slots. Since all defendants prefer* $(a, 3)$ *to* $(l, 3)$ *for any* $l \in \{b, c\}$, *they are all assigned to lawyer a at the end of this step.*

Because lawyer a reaches full capacity in step 1, his capacity remains unchanged in step 2, while the capacities of lawyers b and c are each reduced by one. Given that defendants 1 *and* 2 *prefer* $(b, 2)$ *to* $(a, 3)$, *they are assigned to lawyer b, while defendant* 3 *remains with lawyer a at the end of step 2.*

It is important to note that for any reasonable mechanism ψ, *the outcomes in steps 1 and 2 must align with those produced by serial dictatorship since all defendants are matched with their most preferred lawyers at these steps.*

Step 3 begins with a capacity profile of $(2, 2, 1)$*, reflecting the reduction of one unit in the capacities of lawyers* a *and* c*. At this stage, defendants* 1 *and* 2 *prefer* $(a, 2)$ *to* $(b, 2)$*, and thus, all defendants strictly prefer lawyer* a *over others. Under the serial dictatorship mechanism, defendants* 1 *and* 2 *are assigned to lawyer* a*, while defendant* 3 *is assigned to lawyer* b*.*

In the final step, the capacity profile is $(2, 1, 1)$*. The assignment produced by the serial dictatorship mechanism remains unchanged from the previous step and it is not mobility-free as defendant* 3 *is mobile toward lawyer* a*.* ♦

A closer examination of Example 5 reveals why the properties we imposed thus far do not guarantee a mobility-free outcome for a congestion-cutting algorithm. This failure is due to the fact that, although lawyers' capacities weakly decrease as the algorithm progresses, a defendant may be worse off from one step to another, e.g., defendant 3 is worse off at the end of step 3 compared to step 2. In fact, under her preference $\succ_3$, she strictly prefers her initial assignment at step 1 to her final assignment at step 4. To prevent such failures, we next introduce an additional familiar property of a fixed-capacity mechanism ψ that, when combined with non-wastefulness, ensures mobility-freeness in the final assignment.

Definition 11. *A fixed-capacity mechanism* ψ *is individually rational if no defendant is worse off relative to her initial endowment, i.e., for all* $[\omega, q, P]$ *and for all* $i \in N$*, we have either* $\psi_i[\omega, q, P] \; P_i \; \omega_i$ *or* $\psi_i[\omega, q, P] = \omega_i$*.*

If we fix a non-wasteful and individually rational mechanism ψ for the problem in Example 5, the algorithm differs only in steps 3 and 4. Defendant 3 is guaranteed a slot with lawyer a, and one of defendants 1 or 2 will fill the other slot. Thus, either defendants 1 and 3 or 2 and 3 are assigned to lawyer a, while the remaining defendant from $\{1, 2\}$ is assigned to lawyer b, depending on the choice of ψ. In both cases, the resulting assignment is mobility-free.

Now consider a mechanism ψ that is both non-wasteful and individually rational. Recall that at the start of each step k, defendants retain their lawyer assignments from step $k - 1$, and lawyer capacities weakly decrease over time. Individual rationality of any fixed-capacity mechanism then ensures that no defendant becomes worse off as the algorithm progresses. Since any lawyer with a free slot at a given step serves as an "outside option," a defendant who could have benefited from switching in the final assignment must have had both the opportunity and incentive to do so earlier. That

she did not switch, and weakly improved over time, implies the final assignment is mobility-free. If ψ is also efficient, the final assignment is not only mobility-free but also constrained efficient, and hence efficient by Lemma 1.

This welfare-monotonicity also highlights the main methodological difference between our approach and the *Deferred Acceptance with Voluntary Withdrawals* mechanism of Copland [2023], which builds on the well-known deferred acceptance (DA) algorithm introduced by Gale and Shapley [1962]. In Copland's mechanism, the DA logic is adapted to a school-choice environment with class-size externalities: as the process unfolds, class sizes weakly increase and students move weakly down their feasible preference lists. Congestion-cutting reverses this direction. It begins from maximal hypothetical congestion and reduces the capacities of underdemanded lawyers; when the fixed-capacity mechanism used at each stage is individually rational, the tentative assignment from the previous step is treated as an endowment, and hence every defendant weakly improves from one step to the next. This increasing-welfare feature is central to the proof of Theorem 3. Moreover, the congestion-cutting construction is modular: it is not tied to DA, and can be implemented with any fixed-capacity mechanism satisfying the required properties. Thus, by choosing the stage mechanism appropriately, the designer obtains mobility-freeness, efficiency, and the equilibrium guarantees studied below.

Theorem 3. *If ψ is non-wasteful and individually rational, CC^{ψ} is mobility-free. If, in addition, ψ is efficient, then CC^{ψ} is also efficient.*

Recall from the previous section that an efficient mobility-free mechanism may admit an equilibrium outcome that is not mobility-free. Indeed, a closer look at the mechanism in Example 7, provided in Appendix B, reveals that the outcomes corresponding to the preference profiles in the example can be reached by some CC^{ψ}, where ψ is individually rational and efficient. Therefore, not all mobility-free and efficient congestion-cutting mechanisms ensure mobility-free outcomes in equilibrium. Below, we provide a sufficient condition that identifies a class of congestion-cutting mechanisms with desirable equilibrium outcomes.

Definition 12. *A fixed-capacity mechanism is **group strategy-proof** if no group of agents can gain by jointly misrepresenting their preferences, i.e., for any $S \subseteq N$, for any $i \in S$, any P, and any $\bar{P}_S$, $\psi_i(\bar{P}_S, P_{-S})\ P_i\ \psi_i(P)$ implies that $\psi_j(P)\ P_j\ \psi_j(\bar{P}_S, P_{-S})$ for some $j \in S$.*

Proposition 4. *If ψ is non-wasteful, individually rational, and group strategy-proof, all equilibrium outcomes of the preference revelation game induced by CC^{ψ}, whenever they exist, are mobility-free regardless of the composition of naive and sophisticated defendants, i.e., for any $K \subseteq N$, and for any preference profile $\succ$, all the Nash equilibria of the game $\Gamma^{CC^{\psi}}(K, \succ)$ are mobility-free under $\succ$.*

Theorem 3 enables the construction of various mobility-free mechanisms, provided that the underlying fixed-capacity mechanism is both non-wasteful and individually rational. As a concrete example and a potentially useful practical congestion-cutting mechanism, we adopt the well-known Top Trading Cycles (TTC) mechanism of Shapley and Scarf [1974]. In particular, we use the You-Request-My-House-I-Get-Your-Turn (YRMH-IGYT) procedure of Abdulkadiroğlu and Sönmez [1999], which adapts TTC to house-allocation problems with existing tenants. To apply this procedure to our fixed-capacity problem, we view each lawyer slot as a house: each defendant initially occupies one slot at her endowed lawyer, and any slots not initially occupied by a defendant are vacant.

For consistency with the terminology used above, we refer to the resulting fixed-capacity mechanism as TTC^{π}, where π is a fixed priority order over defendants. Thus, at the outset, defendant i is treated as the existing tenant of one slot at lawyer ω_i, while all other slots are treated as vacant houses. The priority order π initializes the current queue and is also used to break ties among occupied slots of the same lawyer: whenever a requested lawyer has no vacant slot, the defendant with highest priority according to π among those who occupy a slot at that lawyer receives the next turn.

Algorithm 2 (**The top trading cycles (TTC^{π})**). *Fix a strict priority ordering of defendants π and a fixed-capacity problem $[\omega, q, P]$.*

Consider the defendants one-by-one according to a current queue of yet unassigned defendants, which is initially given by the priority order π. Initially, each defendant i occupies one slot at lawyer ω_i, and all other slots are vacant. The current queue determines whose turn it is to demand a lawyer, while the fixed order π is used to break ties among occupied slots of the demanded lawyer. At any moment during the procedure, a lawyer l is called available unless all the q_l slots of the lawyer have already been assigned to defendants by the procedure.

The first defendant in the current queue demands a slot at her favorite available lawyer. If there is a vacant slot at this lawyer, she is assigned to one such slot and

removed together with the assigned slot, while the slot she used to occupy becomes vacant.[38] *If all slots at this lawyer are occupied, then among the defendants who occupy a slot at that lawyer, the one with highest priority according to π is moved to the top of the current queue and is asked to demand a slot at her favorite available lawyer.*

Continuing in this fashion, the defendant at the top of the current queue is either assigned to a vacant slot or gives her turn to the highest-priority defendant according to π who occupies a slot at the demanded lawyer. If the process does not lead to a vacant slot being assigned, then it necessarily leads to a cycle of defendants $(i_1, i_2, \ldots, i_t)$, with $t \geq 1$, such that i_1 demands the lawyer of the slot occupied by i_2, ..., and i_t demands the lawyer of the slot occupied by i_1.[39] *In such a case, assign each defendant in the cycle to the slot occupied by the next defendant in the cycle, and remove these defendants together with the assigned slots. After each assignment and removal, continue with the next defendant in the current queue. The algorithm terminates when all defendants are removed.* ♦

The algorithm above is precisely the YRMH-IGYT procedure applied to the associated slot economy: lawyer slots play the role of houses, defendant i is the existing tenant of one slot at lawyer ω_i, and all other slots are vacant houses. The only additional convention needed in our fixed-capacity setting is the fixed, payoff-irrelevant tie-breaking rule among vacant slots of the same lawyer, since defendants care only about the lawyer they receive. YRMH-IGYT rules belong to the class of hierarchical exchange rules introduced by Pápai [2000].[40] Since hierarchical exchange rules are group strategy-proof, the associated slot mechanism is group strategy-proof. Because defendants care only about the lawyer they receive and the convention among slots of the same lawyer is fixed and payoff-irrelevant, this property carries over to the induced fixed-capacity mechanism TTC^{π} over lawyers. The YRMH-IGYT procedure also yields assignments that are individually rational and efficient; as noted above, efficiency implies non-wastefulness in the fixed-capacity problem.

Corollary 2. *$CC^{TTC^{\pi}}$ is mobility-free and efficient, and it yields the unique envy-free and efficient assignment whenever such an assignment exists. Moreover, it is not obviously manipulable and it only yields mobility-free assignments in equilibrium, whenever one exists, regardless of the composition of naive and sophisticated defendants.*

[38] If several vacant slots of the requested lawyer are available, one of them is selected according to an arbitrary fixed tie-breaking rule.

[39] Note that if $t = 1$, then the defendant forms a "self-cycle" and keeps her slot before being removed.

[40] See, for example, Sönmez and Ünver [2010].

5.2 Practical Implementation

While our theoretical framework enables the design of algorithms that control congestion under optimal assignments, it faces practical challenges when applied in full-scale real-world implementations. Two closely related obstacles are: (i) the impracticality of using the preference-reporting language from our theoretical model *as is*, and (ii) the difficulty of implementing a static mechanism in a dynamic setting with strict legal time constraints, such as the requirement in Texas to assign counsel within 72 hours. In what follows, we address these challenges directly and propose a simple, practical deployment plan to overcome them.

Before and after the Client Choice pilot program, defendants were assigned lawyers through a "rolling roster" system—a common practice not only in the United States but also globally. The Client Choice program replaced this forced assignment model with a system based on defendant preference, but at the cost of unintended variance in lawyers' congestion. Our proposed implementation builds on a carefully designed hybrid of these two approaches, preserving choice while managing congestion effectively in a dynamic assignment environment.

In our model, defendants are asked to rank $m \times n$ pairs, where each pair consists of a lawyer and a congestion level. During the 2015 Client Choice pilot program, 795 defendants and 54 lawyers participated over a period of slightly more than a year, with participants entering and exiting the program at different times. These dynamics—particularly the rolling nature of entry and exit and the cognitive burden of ranking prohibitively many options—make it necessary to structure the system as a series of smaller, more manageable fragments in which defendants are processed in **batches**.

Imagine initiating the process at time $t = 0$, when we begin admitting incoming defendants into the first batch, according to their arrival times. The composition of this batch is finalized whenever the earlier of the following two events occurs: (1) a target number T of defendants is reached, or (2) a fixed amount of time (e.g., 48 hours) has passed since the first defendant was admitted.

Meanwhile, all participating lawyers are placed on a rolling roster. Defendants in the batch are each presented with a small set of lawyers to choose from. To mimic the approach used in Comal County, for instance, we would select the first three lawyers from the wheel for the first batch.

Once a defendant has been given time to review basic information about the three lawyers (e.g., via short booklets or profiles), they are shown a simple schematic pref-

erence table. This table consists of three vertical bars—one per lawyer—each divided into color-coded regions: green (low congestion), yellow (moderate congestion), and red (high congestion). Each color corresponds to an estimated level of the lawyer's workload at the time of assignment.

The defendant is then asked to indicate preference by placing numbers 1 through 9—with '1' representing highest preference, '2' second-highest preference, etc.—on all combinations of the form (lawyer, congestion level) based on the following illustrative interpretations:

Green zone: *Your lawyer will have very few other clients and is highly likely to be responsive to your needs.*

Yellow zone: *Your lawyer will have a moderate number of other clients and is reasonably likely to be responsive, though occasional delays may occur.*

Red zone: *Your lawyer will have many other clients and is likely to be slow or less attentive to your case due to their workload.*

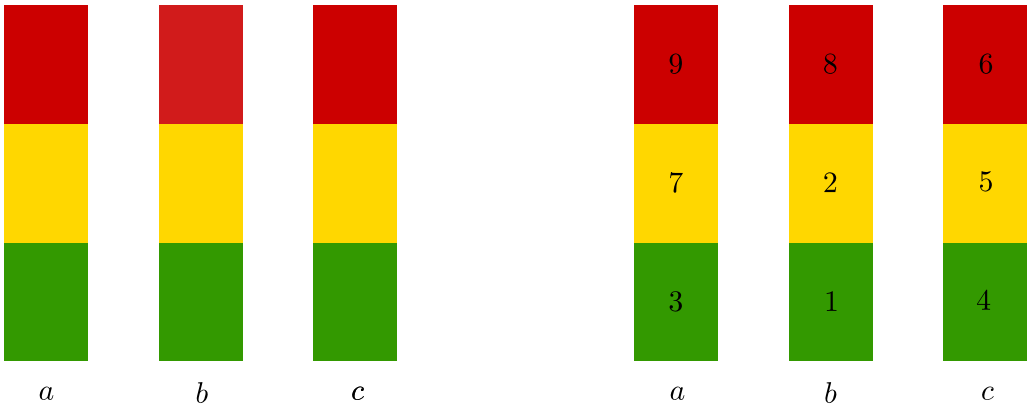


Figure 4: A schematic preference table provided to defendants in a 3-lawyer batch (left). An example of a filled-out preference table indicating first preference for lawyer b with low congestion, second preference for lawyer b with moderate congestion, followed by third preference for lawyer a with low congestion, and so on (right).

We now turn to explain how the color-coded preference bars shown to each batch of defendants are constructed. The height of each colored region corresponds to the maximum number of defendants for that congestion level. For a given bar representing some lawyer l, let l_g and l_y denote the heights for the green and yellow areas of lawyer l. In Comal County, the lawyers who are deemed qualified to participate in the Client Choice pilot program were determined by the Criminal and Misdemeanor Court judges.

In line with this, we imagine that these administering judges choose the parameters l_g and l_y individually for each lawyer l based on the lawyer's profile and idiosyncratic characteristics such as experience, specialty, and availability. These parameters may also be based on a lawyer's self-declared congestion zones, subject to the discretion of the overseeing judicial body.

Our design specifically aims to guarantee each defendant *red-zone avoidance*: the option to avoid a highly congested lawyer regardless of other defendants' preferences, provided that her preferences reflect a clear desire for low congestion, i.e., she ranks red-zone options only after all available green- and yellow-zone options. Since we consider only mobility-free mechanisms, ensuring red-zone avoidance is feasible by setting the batch size T equal to the sum of non-red-zone capacities across all lawyers in the batch: $T := \sum_{l \in B} l_y$, where B denotes the set of lawyers in a batch.

In order to be consistent with our baseline model, the above implementation leaves the size of a red zone unrestricted. It is, however, straightforward to adapt the practical implementation to lawyer-specific capacity caps.[41] Importantly, the red-zone avoidance guarantee continues to hold in this extension.

Given a defendant preference indicated through ranks assigned on the color-coded bars, the transformation to preferences over lawyer-congestion level pairs as required by our model is fairly straightforward: Take any color-coded preferences for a given defendant i. The first choice of the defendant is always a lawyer, say l, with green zone capped at l_g. This implies $(l,1) \succ_i (l,2) \succ_i \cdots \succ_i (l,l_g)$ as the top l_g alternatives. If the second choice is the yellow zone of the same lawyer, we extend the ranking up to the pair (l,l_y). If the second choice is some other lawyer, say l', with the green zone capped at l'_g, then this implies the portion of preferences $(l',1) \succ_i (l',2) \succ_i \cdots \succ_i (l',l'_g)$. Continuing in this fashion, we follow the rankings to transform color-coded preferences into full-fledged preference ordering over lawyer–congestion level pairs.

When a lawyer finishes all the cases he was assigned, he is added back into the rotating wheel. In case of excess demand for lawyer services, our design also makes it possible to feed a lawyer back into the wheel before exhausting all his cases. However, this needs to be done with extra care in order not to violate the congestion preferences of his existing clients. For an illustration, consider some lawyer l who is initially assigned to many cases that would fill his caseload up to the upper bound of his yellow area.

[41] All that is necessary to adopt a congestion-cutting algorithm is to set the capacity of each lawyer equal to its cap (rather than n) at the initialization of the algorithm. The definition of mobility-freeness must also be adapted to account only for deviations that are feasible given lawyer caps.

Suppose that, over time, the congestion level of lawyer l falls to the green area as he dispenses with his cases. If one wishes to make this lawyer available for new clients, the color-coded preferences for him should now only include the green and yellow areas, precluding high congestion levels that would hurt existing clients.

A challenge that emerged during the Client Choice pilot in Comal County was the concentration of client selections among a small number of lawyers. As reported in the evaluation, "there were consistently a small number of lawyers, between three and five, who represented about a third of all Client Choice cases." These lawyers continued to attract clients over multiple months, reinforcing their visibility and increasing the chances of being selected again. According to the report, "defendants primarily cited the lawyer's reputation, followed by years of experience, recommendation from a family member or friend, and prior representation of the defendant by the lawyer" as key factors influencing their decisions. Some lawyers, by contrast, were not chosen at all and eventually stopped participating in the program. Given that reputation and experience were the most commonly cited factors, lesser-known or less experienced lawyers were likely at a disadvantage—further amplifying the gap between a small number of frequently selected attorneys and the rest. This dynamic raises concerns about the long-term sustainability of the system: if the number of active participating lawyers declines and demand continues to concentrate on a few individuals, congestion becomes difficult to manage.

During JMI's baseline site visit in 2013, several participating lawyers expressed motivations for accepting indigent defense cases that went beyond financial incentives. Despite a general view that the compensation was well below private market rates, some lawyers cited the opportunity to gain courtroom experience, a sense of duty to give back to the community, and a desire to maintain a steady supplement to their private practices. These motivations suggest that younger or lesser-known lawyers might remain engaged if given a meaningful opportunity to receive assignments and build a track record.

Our batch-based approach offers a natural way to address this issue. For instance, some batches could be entirely composed of less experienced lawyers, which would give them a fair opportunity to gain exposure and experience. While defendants in such batches might initially face more limited choices, the longer-term effect is likely to be beneficial for the system overall, by increasing the number of experienced lawyers available in future periods. Additionally, if a lawyer receives very few or no selections in a mixed-experience batch, he may be included as the third option in subsequent

batches. This gradual adjustment contributes to the long-term sustainability of the program by supporting the development and retention of a broad and capable group of participating lawyers.

Finally, we note that, in practice, an even simpler and more automated preference reporting language can be employed—one that allows defendants to delegate the full construction of their preferences to the system. In this alternative model, preferences over lawyers and their congestion levels are elicited separately. First, the defendant ranks the three lawyers assigned to their batch, similar to the approach used in Comal County. Then, they indicate how much they care about their lawyer's availability or workload relative to the ranking they just provided, by selecting from a menu of options.[42] Using an additive or multiplicative scoring rule based on the defendant's responses, the system can then automatically generate a complete ranking over all lawyer–congestion level combinations.

6 Literature review

The lawyer assignment problem has received little attention from economists. To our knowledge, Spurr [1987] is the first to consider allocating lawyers to cases; however, he focuses on a market mechanism with transferable utility. Kamali Shahdadi [2018] provides a stylized moral-hazard model studying the connection between assigning lawyers to indigent defendants and lawyers' incentives to exert effort. Aköz et al. [2025] study a generalization of the pairwise stability concept of Gale and Shapley [1962] to the market for legal services; however, their focus is a four-sided matching model resembling civil rather than criminal litigation. Iossa and Jullien [2012] study a simple model of civil litigation and analyze how the market allocation of lawyers is affected by career concerns of judges. Haruvy et al. [2006] consider the matching of judicial clerks to judges. Agan et al. [2021], Shem-Tov [2022], Iyengar [2007], Anderson and Heaton [2012] empirically

[42]For example, this can take the form of a multiple-choice question such as: *How much do you care about your lawyer's time and attention to your case, given that they may also be assisting other defendants? A: VERY MUCH – I'd prefer my second or third choice if my top-choice lawyer is very busy and may be less attentive to me. B: MODERATELY – I'd still prefer my top choice unless they are much busier than the others. C: SLIGHTLY – I generally prefer my top choice, but if they're extremely overloaded, I might consider a less busy alternative. D: DON'T CARE – I want my first-choice lawyer no matter how busy they are, and I understand they may not be as available as I want them to be.* Whereas option A would imply congestion-lexicographic preferences, option D would imply lawyer-lexicographic preferences. Options B and C would correspond to intermediate types of preferences.

study the assignment of lawyers to indigent defendants and compare the performance of lawyers serving as public defenders to that of lawyers with private practices. While legal scholars have extensively discussed the role of defendants' choice [Brecher, 1987, Nugent-Borakove and Cruz, 2017, Garrett, 2008, Tague, 1974, Schulhofer and Friedman, 1993, Rappaport, 2017, Garcia, 1991, Ogletree Jr, 1992, Baker, 2017], the focus is on individual assignments rather than the overall allocation of defendants to lawyers. To our knowledge, no prior study considers congestion in this context.

On the game-theoretic side, our model belongs to the literature on congestion games and games with partial rivalry. Rosenthal [1973] showed that classical congestion games admit a pure-strategy Nash equilibrium by constructing an exact potential, and Monderer and Shapley [1996] later placed this insight in the broader framework of potential games and their better-reply dynamics. The setting closest to ours is the one in which each player chooses a single resource from a common finite set and her payoff depends on the resource and the number of other players who choose the same resource. Unlike in Rosenthal [1973] and Monderer and Shapley [1996], and similar to our model, Milchtaich [1996] allows for player-specific payoffs. In this setting, Milchtaich [1996] proves the existence of a pure-strategy Nash equilibrium; in our lawyer-assignment problem, this yields the existence of a mobility-free allocation. Moreover, Milchtaich's constructive argument provides a best-reply path from any initial assignment to equilibrium, which in our finite setting gives a polynomial-time way to find a mobility-free allocation. Quint and Shubik [1994] independently study a related migration model in which animals choose territories and payoffs depend on location and the number of co-located animals, again obtaining pure-strategy equilibrium existence. Finally, Konishi et al. [1997] show that under anonymity, independence of irrelevant choices, and partial rivalry, strong Nash equilibria exist; in our setting, this implies the existence of efficient and mobility-free assignments. These papers adopt a non-cooperative equilibrium perspective, whereas our paper studies a direct mechanism-design problem motivated by lawyer assignment.

From a normative assignment perspective, Bogomolnaia and Moulin [2025] introduce two fairness notions for assignment problems under congestion preferences: top-n fairness and competitiveness. They primarily study the feasibility of these notions in deterministic and probabilistic settings, and show that the latter is equivalent to the conjunction of efficiency and envy-freeness. Efficient and mobility-free assignments in our setting are always top-n fair, and coincide with competitive allocations whenever such allocations exist. Bogomolnaia and Moulin [2025] also note that the equilibrium

of any top-n fair mechanism—and hence of any mobility-free mechanism—guarantees a top-n outcome for every agent. We complement this observation by providing direct mechanisms that further guarantee mobility-freeness in equilibrium, even when some agents are naive. Chen et al. [2023] provide a polynomial-time algorithm for finding competitive allocations whenever they exist.

In two related papers, Phan et al. [2024] and Copland [2023] study negative externalities due to crowding in the context of school choice.[43] Phan et al. [2024] model congestion arising from the shared use of a continuous resource, such as school funding, which may be subject to inefficiency or waste. They propose a novel solution concept called *rationing crowding equilibrium* (RCE), in which a school's resource level plays a role analogous to price in a classical Walrasian equilibrium. Our work and that of Phan et al. [2024] differ substantially in concept, modeling, and focus. Although a lawyer's time endowment can be viewed as a continuous resource, courts can only assign cases to lawyers, but lack both the ability and the authority to mandate workload adjustments aimed at eliminating envy among defendants. Consequently, we model the lawyer assignment problem as a discrete allocation problem, making their approach inapplicable. Moreover, while we rely on ordinal preferences, their algorithm assumes linear preferences, where a student's payoff equals the product of an idiosyncratic school value and her share of resources. Such an assumption is not entirely realistic in legal contexts, where the effects of congestion are typically nonlinear and context-dependent.[44]

Although an approximation of RCE could potentially be applied to problems with discrete congestion, in line with the school choice literature, the focus of Phan et al. [2024] is eliminating justified envy while maintaining strategy-proofness. In fact, Phan et al. [2024] demonstrate that eliminating justified envy requires wasting some resources. While such inefficiencies may be negligible for problems involving a large number of agents on both sides of the market (e.g., school choice), it is problematic for smaller environments like ours. In contrast, given the generally heavy workload put on public lawyers, our primary goal is achieving efficiency alongside mobility-freeness—and, when feasible, envy-freeness. On the other hand, our result on the bounds on manipulation suggests that whenever envy-freeness is feasible, defendants cannot manipulate a mobility-free mechanism.

[43]Similar in spirit to Phan et al. [2024], Tierney [2019] studies congestion in an environmental resource allocation problem.

[44]In particular, the congestion levels are especially important in the neighborhood of caseload limits of 2-8 cases per week [New York State Office of Indigent Legal Services, 2014, Lefstein, 2011].

Copland [2023] focuses on achieving a fair school choice assignment similar to Phan et al. [2024], while as in our approach, he works in a discrete setting. He introduces a fairness notion combining no-justified-envy and mobility-freeness, and develops a variant of the deferred acceptance algorithm, called *Deferred Acceptance with Voluntary Withdrawals*, which achieves a fair outcome. By contrast, our approach emphasizes not only fairness, but also efficiency and desirable incentive properties. While these studies concentrate on school choice, we introduce lawyer assignment under congestion as a novel market design application.[45]

7 Conclusion

This paper began by highlighting a pressing issue in the American criminal justice system: the appointment of public defenders to indigent defendants without any consideration of their preferences or the workload of the assigned attorneys. While the current rolling roster system is administratively simple and appears fair on its face, it fails to account for differences in attorney quality, specialization, and capacity—ultimately generating inefficiencies and perceptions of procedural injustice.

The Client Choice pilot program in Comal County, Texas, offered a promising alternative by allowing defendants to choose their attorneys. While the program demonstrated several positive outcomes—including increased engagement between lawyers and clients, and improved plea and trial outcomes—it also confirmed a major theoretical and practical challenge: unconstrained choice in the presence of limited lawyer capacity can lead to severe congestion and inequitable case distribution.

To address these concerns, we proposed a novel class of assignment mechanisms—congestion-cutting algorithms—that directly incorporate defendants' preferences over both lawyers and their caseloads. Our mechanism design framework emphasizes two key normative criteria: efficiency and mobility-freeness, ensuring that no defendant can be made better off without harming another and that no one can unilaterally improve their outcome by switching lawyers. Unlike the Comal mechanism, our approach guarantees the existence of desirable allocations and avoids obvious manipulability.

[45]Insights from economic theory have been successfully applied to various areas, including organ transplants [Roth et al., 2004], residency matching [Roth and Peranson, 1999], doctor allocation [Aygün and Bilgin, 2023] and school choice [Abdulkadiroğlu and Sönmez, 2003]. While Aygün and Bilgin [2023] consider adjusting doctor allocation algorithms for flexible quotas, they do not explicitly model congestion.

Beyond the formal guarantees, congestion-cutting mechanisms offer a realistic and implementable path forward for jurisdictions considering a shift toward a choice-based legal aid system. These mechanisms preserve the core insight of the Client Choice program—that giving defendants a voice in their representation can yield significant benefits—while addressing its core flaw: failing to manage lawyer congestion in a principled and systematic way.

The broader implication is clear: choice and fairness need not be at odds with administrative feasibility. By carefully designing mechanisms that respect both the preferences of defendants and the resource constraints of the public defense system, we can make meaningful progress toward a more just, efficient, and humane system of legal aid in the United States.

APPENDIX

A Proofs

Proof of Proposition 1 Fix $\succ$ and let μ be the unique envy-free and efficient assignment under $\succ$. Then μ is mobility-free and efficient. Suppose for a contradiction $\mu' \neq \mu$ is also mobility-free and efficient. As no agent envies another under the congestion profile s^{μ}–and hence, all defendants get their best option–μ is the only efficient assignment with congestion profile s^{μ}. Therefore, μ' has a different congestion profile. Let $\bar{L}$ be the set of lawyers that have strictly fewer clients in μ' compared to μ. Then, at least one defendant i, who is assigned a lawyer $\bar{l} \in \bar{L}$ at μ, is assigned at μ' a lawyer $l \in L \setminus \bar{L}$ where $s_l^{\mu} \leq s_l^{\mu'}$. If $\mu_i' \succ_i \mu_i$, i would have envied the defendants assigned to l at μ in case $s_l^{\mu} > 0$, contradicting that μ is envy-free, and i would have been willing to move to l from her assignment at μ in case $s_l^{\mu} = 0$, contradicting that μ is efficient. If $\mu_i \succ_i \mu_i'$, i would be mobile toward $\bar{l}$ at μ' as $s_{\bar{l}}^{\mu'} < s_{\bar{l}}^{\mu}$, contradicting that μ' is mobility-free. ■

Proof of Proposition 2 Fix N, L with $m \geq 2$, and a set of naive agents $K \neq \emptyset$. Now, consider the following preference profile. All $j \in N \setminus K$ are lawyer–a–lexicographic, i.e., the top n alternatives in $\succ_j$ are ordered as follows: $(a,1) \succ_j (a,2) \succ_j \cdots \succ_j (a,n)$. Therefore, it is a dominant strategy to announce $(a,1)$ under the game $\Gamma^{CM}(K, \succ)$ for all $j \in N \setminus K$. Now let $(a,1) \succ_i (b,1) \succ_i \cdots$ for all $i \in K$, where $(b,1)$ is the second best alternative in $\succ_i$. Since $\succ_i$ is the only strategy for agent $i \in K$, all agents are assigned to a at the unique Nash equilibrium outcome assignment of the game $\Gamma^{CM}(K, \succ)$. However, all agents in K are mobile toward b under $\succ$ at this assignment. ■

Proof of Proposition 3 Fix a mobility-free mechanism φ, a defendant $i \in N$, and a true preference $\succ_i$ for i.

First, to show that the best-case scenario for agent i cannot be improved, it suffices to show that the top alternative in $\succ_i$ is always achieved at φ for some $\succ_{-i}$. To be precise, note that the top alternative for agent i is always $(l,1)$ for some $l \in L$, and let it be $(a,1)$. Now consider the case where all other defendants have lawyer–b–lexicographic preferences, i.e., $\succ_{-i}$ is such that for all $j \neq i$ and for all $l \neq b$, $(b,n) \succ_j (l,1)$. Then the unique mobility-free assignment under $\succ$ is μ with $\mu_i = (a,1)$ and $\mu_j = (b, n-1)$ for all $j \neq i$. Thus, $\varphi_i(\succ_i, \succ_{-i}) = (a,1)$.

We now discuss that i cannot improve the worst-case scenario given her true pref-

erence $\succ_i$. Let $x = (l, r)$ for some $l \in L$ and r be the nth-ranked alternative in $\succ_i$. Bogomolnaia and Moulin [2025] show that in any mobility-free assignment, i's assignment is at worst the nth-ranked alternative in her *reported preference*. This implies that either $\underline{\mu}_i^{\varphi}(\succ_i; \succ_i) = x$ or $\underline{\mu}_i^{\varphi}(\succ_i; \succ_i) \succ_i x$. Now, consider any other reported preference $\succ_i' \neq \succ_i$. To show that i cannot improve her worst-case scenario outcome by reporting $\succ_i'$, it suffices to show that either $x = \underline{\mu}_i^{\varphi}(\succ_i'; \succ_i)$ or $x \succ_i \underline{\mu}_i^{\varphi}(\succ_i'; \succ_i)$; that is, the worst-case scenario according to i's true preference $\succ_i$ when i reports $\succ_i'$ is either x or worse.

Below, we first show that for any reported preference $\hat{\succ}_i$, any one of the top-n alternatives in $\hat{\succ}_i$ is achieved as the unique mobility-free outcome for defendant i at some profile for an appropriate selection of others' preferences $(\succ_{-i})$.

Claim 1. For any reported preference $\hat{\succ}_i$ and for any $k \leq n$, there exists a profile $\succ_{-i}$ of others such that *in any mobility-free assignment under* $(\hat{\succ}_i, \succ_{-i})$, i obtains the kth-ranked alternative with respect to $\hat{\succ}_i$.

Proof of Claim 1. Fix some $k \leq n$. Without loss of generality, let i be agent 1 and let (a, s) be the kth-ranked alternative in $\hat{\succ}_1$. Now we construct the profile $\succ_{-1}$ for the rest of the agents. For each $j \in \{2, ..., s\}$, let $\succ_j$ be lawyer–a–lexicographic, as above; that is, $(a, n) \succ_j (l, 1)$ for all $l \neq a$. This ensures that all agents in $\{2, ..., s\}$ are assigned to lawyer a in any mobility-free assignment under $(\hat{\succ}_1, \succ_{-1})$.

To finish the argument, we need to construct the rest of the profile for the defendants in $\{s+1, ..., n\}$ such that: (i) defendant 1 is assigned to lawyer a in any mobility-free assignment, and (ii) no defendant in $\{s+1, ..., n\}$ is assigned to a in any mobility-free assignment, and hence, defendant i's outcome is (a, s), given that all defendants in $\{2, ..., s-1, s\}$ are assigned to lawyer a. Since (a, s) is the kth-ranked alternative, by Assumption 1, we have either $s = k$ or $s < k$.

Case 1: $s = k$. By Assumption 1, we have $(a, 1)\hat{\succ}_1(a, 2)\hat{\succ}_1 \cdots \hat{\succ}_1(a, s-1)\hat{\succ}_1(a, s)$. Now, fix an arbitrary lawyer $b \neq a$, and for all $j \in \{s+1, ..., n\}$ set $\succ_j$ such that $(b, n) \succ_j (l, 1)$ for all $l \neq b$. Then, in the unique mobility-free assignment, all defendants in $\{1, ..., s\}$ are assigned to a, and the rest are assigned to b.

Case 2: $s < k$. Let $s' = k - s > 0$. Define $L' \subseteq L \setminus \{a\}$ as the set of lawyers that appear in the top-k alternatives in $\hat{\succ}_1$, and for all $l \in L'$, let $\bar{s}_l$ be the highest congestion level for which $(l, \bar{s}_l)$ appears among the top-k alternatives in $\hat{\succ}_1$. Now, we define the preferences of s' defendants in $\{s+1, s+2, ..., k\}$. It is straightforward to verify that $\sum_{l \in L'} \bar{s}_l = s'$ by Assumption 1. Then we can partition these s' defendants in a way

that for each $l \in L'$, exactly $\bar{s}_l$ defendants from $\{s+1, ..., k\}$ lexicographically prefer l to any other lawyer regardless of the level of congestion at l. This ensures that: (iii) for any $l \in L'$, at least $\bar{s}_l$ defendants are assigned to l in any mobility-free assignment, and (iv) no defendant in $\{s+1, ..., k\}$ is assigned to a in any mobility-free assignment. Finally, fix an arbitrary lawyer $b \in L'$, and for all $j \in \{k+1, ..., n\}$ set $\succ_j$ such that $(b, n) \succ_j (l, 1)$ for all $l \neq b$. Then, in any mobility-free assignment all defendants in $\{k+1, ..., n\}$ are assigned to $b \neq a$, and therefore by (iv), (ii) holds. It remains to show that defendant 1 is assigned to lawyer a in any mobility-free assignment. Suppose not. Since exactly $s-1$ defendants other than defendant 1 are assigned to lawyer a, defendant 1 would obtain (a, s) by moving to lawyer a. If defendant 1 is assigned to some $l \in L'$, then by (iii), the congestion level at l is at least $\bar{s}_l + 1$ once defendant 1 is also assigned to l. By the definition of $\bar{s}_l$ and Assumption 1, her assignment is worse than (a, s) according to $\hat{\succ}_1$. If defendant 1 is assigned to some lawyer outside $L' \cup \{a\}$, then even that lawyer at congestion level 1 is outside the top-k alternatives in $\hat{\succ}_1$, and hence her assignment is again worse than (a, s). In either case, defendant 1 is mobile toward lawyer a, a contradiction. Therefore, defendant 1 is assigned to lawyer a in every mobility-free assignment, and since exactly $s-1$ other defendants are assigned to a, her outcome is (a, s). □

Together with the nth-ranked alternative upper bound established by Bogomolnaia and Moulin [2025], Claim 1 implies that the worst-case scenario outcome for i given her true preference $\succ_i$ under φ when she reports $\succ_i$ is $\underline{\mu}_i^{\varphi}(\succ_i; \succ_i) = x$. If $\succ_i'$ has the same top-n set as $\succ_i$, possibly in a different order and with the remaining alternatives ordered arbitrarily, by Claim 1, each one of the top-n alternatives in $\succ_i'$ (and hence, in $\succ_i$) is achieved as the unique mobility-free outcome of i at some profile $(\succ_i', \succ_{-i})$. Then, the nth-ranked alternative upper bound established by Bogomolnaia and Moulin [2025] implies $x = \underline{\mu}_i^{\varphi}(\succ_i'; \succ_i)$. Now, consider the only alternative case where there is $y = (l', r')$ such that y is one of the top-n alternatives in $\succ_i'$ but it is not included in the set of top-n alternatives in $\succ_i$. By Claim 1, there is $\succ_{-i}$ such that $\varphi_i(\succ_i', \succ_{-i}) = y$. Therefore, either $x \succ_i y = \underline{\mu}_i^{\varphi}(\succ_i'; \succ_i)$ or $x \succ_i y \succ_i \underline{\mu}_i^{\varphi}(\succ_i'; \succ_i)$. ■

Proof of Theorem 2 (i) Consider the problem in Example 2. It is easy to verify that there are three mobility-free assignments, which are identical up to permuting the defendants. Two defendants are assigned to lawyer a and one of them is assigned to b. Let φ be a mobility-free mechanism and $\mu_i = (b, 1)$. If defendant i announces $\succ_i'$ that she prefers lawyer a to b irrespective of the congestion levels of these lawyers, all

mobility-free allocations still assign two defendants to a, and one to b. In addition, agent i needs to be assigned to a given $\succ_i'$. Therefore, i can manipulate φ at $\succ$.

(ii) Fix any mobility-free mechanism φ, $i \in N$, $\succ$, and $\succ_i'$. Let $\succ'= (\succ_i', \succ_{-i})$, $\varphi(\succ_i', \succ_{-i}) = \mu$, and say without loss of generality $\mu(i) = a$. If μ is mobility-free under $\succ$, the statement in (ii) holds trivially. Now, let μ not be mobility-free under $\succ$. Note that μ is mobility-free under $\succ'$ as φ is a mobility-free mechanism. Then, since $\succ_j=\succ_j'$ for all $j \neq i$, the only mobile agent at μ under $\succ$ is i. Let i be mobile at μ under $\succ$ toward some lawyer $b \in L$, i.e., $(b, s_b^\mu + 1) \succ_i \mu_i$. With the help of the following claim, we first show that when we remove agent i and consider the problem $(N \setminus \{i\}, \succ_{-i})$, there is an assignment $\hat{\mu}$ that is mobility-free under $\succ_{-i}$ such that b's congestion level at $\hat{\mu}$ is not more than s_b^μ. While the claim is stated in general terms, it may be helpful for the reader to interpret it by considering a specific instance: think of $(\bar{N}, \bar{\succ}) = (N \setminus \{i\}, \succ_{-i})$, let $\bar{\mu}$ be the restriction of μ to $N \setminus \{i\}$, and let $\bar{l} = a$. This perspective can aid in understanding the construction presented in the premise of the claim and the way it is applied in the proof.

Claim 2: Let $\bar{N}$ be a set of defendants, $\bar{\mu}$ be an assignment for the problem $(\bar{N}, \bar{\succ})$ where there exists a unique lawyer $\bar{l}$ such that:

(1) any mobile agent (if there is any) is mobile only toward $\bar{l}$ under $\bar{\succ}$, but they would not have been mobile if $\bar{l}$ had one more client,

(2) no agent that is assigned to $\bar{l}$ at $\bar{\mu}$ is mobile and, moreover, none of them would have been mobile even if $\bar{l}$ had one more client.

Then, there is $\hat{\mu}$ which is mobility-free under $\bar{\succ}$ such that:

(*) either $\hat{\mu}$ and $\bar{\mu}$ have the same congestion profile,

(**) or $\bar{l}$ has one more client at $\hat{\mu}$ compared to $\bar{\mu}$, while another lawyer has one less client and all remaining $(m-2)$ lawyers have the same congestion at these two assignments.

Proof of Claim 2: Take any mobile agent, say i_1, and move i_1 to $\bar{l}$, say from $l^1 = \bar{\mu}(i_1)$, while keeping the assignment of all remaining agents. We call the new assignment μ^1. Agent i_1 gets better off with respect to $\bar{\mu}$. At μ^1, no agent is mobile toward $\bar{l}$ by (1), and any agent could only be mobile toward l^1, including those who were initially at $\bar{l}$ by (2) as l^1 is the only lawyer that became more attractive. Since initially (at $\bar{\mu}$) with one more client at l^1 no one was mobile toward l^1, μ^1 satisfies (1) for lawyer l^1. Note that there is one more agent assigned to $\bar{l}$ now and one less assigned to l^1. Since initially agents assigned to l^1 could only be mobile toward $\bar{l}$ at $\bar{\mu}$, none of them is mobile at μ^1

and none of them would have been mobile even if l^1 had one more client. Therefore, μ^1 also satisfies (2) for lawyer l^1. If moreover no one is mobile toward l^1, μ^1 is mobility-free and since μ^1 satisfies (**), Claim 2 holds.

Consider the case that μ^1 is not mobility-free. Move a mobile agent at μ^1, say i_2, from $l^2 = \mu^1(i_2)$ to l^1, and call the new assignment μ^2. If $l^2 = \bar{l}$, μ^2 has the same congestion profile with $\bar{\mu}$ and μ^2 satisfies conditions (1) and (2) of the claim for lawyer l^2. If $l^2 \neq \bar{l}$, the congestion profile of μ^2 is as stated in (**). Moreover, i_2 gets better off with respect to $\bar{\mu}$ and μ^1, and no agent among those who are not initially assigned to lawyer $\bar{l}$ is worse off with respect to $\bar{\mu}$. In this case, since μ^1 satisfies (1) and (2) for l^1, by using the same arguments to verify it for μ^1, μ^2 satisfies conditions (1) and (2) of the claim for lawyer l^2. If μ^2 is mobility-free, Claim 2 holds.

In case μ^2 is not mobility-free, continue moving one mobile agent to the lawyer she is mobile toward, while keeping the assignment of remaining agents the same. For each $k > 2$ and for each μ^k we reach, μ^k satisfies the conditions (1) and (2) for lawyer l^k and the congestion profile of μ^k is either equal to $\bar{\mu}$ or it is as in (**). Therefore, any agent i^k that moves in round k is strictly better off at μ^k compared to μ^{k-1} and never gets worse off in the following rounds. Thus, by the finiteness of the set of agents, this procedure should stop at some k, resulting in a mobility-free assignment that satisfies (*) or (**). □

Now, define $(\bar{N}, \bar{\succ}) = (N \setminus \{i\}, \succ_{-i})$, $\bar{\mu}$ as the restriction of μ to $N \setminus \{i\}$, and $\bar{l} = a$. Since μ is mobility-free at the problem $(N, (\succ_i', \succ_{-i}))$, the assignment $\bar{\mu}$ satisfies conditions (1) and (2) in Claim 2 for the problem $(N \setminus \{i\}, \succ_{-i})$, which is obtained by removing agent i, and lawyer a. Then, Claim 2 immediately implies that there is a mobility-free assignment $\hat{\mu}$ for the problem $(N \setminus \{i\}, \succ_{-i})$ such that either $s_b^{\hat{\mu}} = s_b^{\mu}$ or $s_b^{\hat{\mu}} = s_b^{\mu} - 1$.

The following claim directly proves the statement (ii) in Theorem 2.

Claim 3: There is an assignment $\tilde{\mu}$ that is mobility-free at the problem $(N, \succ)$ such that either $\tilde{\mu}_i = (b, s_b^{\mu} + 1)$ or $\tilde{\mu}_i \succ_i (b, s_b^{\mu} + 1)$.

Proof of Claim 3: Fix the assignment $\hat{\mu}$ that is mobility-free at the problem $(N \setminus \{i\}, \succ_{-i})$ where $s_b^{\hat{\mu}} \leq s_b^{\mu}$. Consider the assignment μ^* for the problem $(N, \succ)$ which differs from $\hat{\mu}$ only by adding agent i to her best lawyer l^* under $\succ_i$ given the fact that she will increase the congestion level at this lawyer by 1 compared to $\hat{\mu}$, i.e., $\mu_i^* = (l^*, s_{l^*}^{\hat{\mu}} + 1) \succ_i (l', s_{l'}^{\hat{\mu}} + 1)$ for all $l' \neq l^*$ and $\mu^*(j) = \hat{\mu}(j)$ for all $j \neq i$. Since $s_b^{\hat{\mu}} \leq s_b^{\mu}$, either $(l^*, s_{l^*}^{\hat{\mu}} + 1) = (b, s_b^{\mu} + 1)$ or $(l^*, s_{l^*}^{\hat{\mu}} + 1) \succ_i (b, s_b^{\mu} + 1)$. If μ^* is mobility-free

under $(N, \succ)$, set $\mu^* = \tilde{\mu}$, and hence the claim holds.

Consider the case that μ^* is not mobility-free under $\succ$. Since $\hat{\mu}$ is mobility-free under $(N \setminus \{i\}, \succ_{-i})$, an agent is mobile at μ^* only if she is assigned to l^*. Pick such an agent i_1 and move her to her best lawyer toward which she is mobile. Note that similar to i, i_1 is assigned to her best lawyer, say l^1, under $\succ_{i_1}$ considering the fact that she will increase the congestion level at this lawyer by 1 compared to $\hat{\mu}$, i.e., $\mu^1_{i_1} = (l^1, s^{\hat{\mu}}_{l^1} + 1) \succ_{i_1} (l', s^{\hat{\mu}}_{l'} + 1)$ for all $l' \neq l^1$ and $\mu^1(j) = \hat{\mu}(j)$ for all $j \neq i_1$. Call this assignment μ^1. If μ^1 is mobility-free, the claim is proved. If not, an agent is mobile at μ^1 only if she is assigned to l^1. Pick such an agent i_2 and move her to her best lawyer toward which she is mobile. Call this assignment μ^2.

Note that in each round, μ^k reached by this procedure is such that the congestion of only one lawyer increases by 1 compared to $\hat{\mu}$ and the rest of the congestion profile is the same. Therefore, an agent who moved in the previous rounds will never move again. Thus, this procedure terminates, resulting in a mobility-free assignment $\tilde{\mu}$ for the problem $(N, \succ)$. Moreover, since l^* never has more clients compared to μ^* in any round and the other lawyers never have fewer clients compared to μ^*, i would never move during this procedure. Therefore, $\tilde{\mu}_i$ is either $(l^*, s^{\hat{\mu}}_{l^*} + 1)$ or $(l^*, s^{\hat{\mu}}_{l^*})$, which is not worse than $(b, s^{\mu}_{b} + 1)$ under $\succ_i$. ■

Proof of Theorem 3 Let ψ be non-wasteful and individually rational. Fix a problem $(N, L, \succ)$ and suppose for a contradiction that $CC^{\psi}(\succ) = \mu$ is not mobility-free under $\succ$. Let i be mobile at μ toward lawyer a under $\succ$, i.e., $(a, s^{\mu}_{a} + 1) \succ_i \mu_i$. Consider the final step k in the CC^{ψ} algorithm such that a has one more capacity compared to the last step, i.e., $q^k_a = q^{k+1}_a + 1 = s^{\mu}_{a} + 1$. Let b be the lawyer i is assigned at the end of step k, and q^k_b be his capacity in this step. Since ψ is individually rational and the capacities of lawyers weakly decrease as the algorithm progresses, i should weakly prefer her final outcome to her tentative outcome at the end of step k under $\succ_i$, i.e., either $\mu_i = (b, q^k_b)$ or $\mu_i \succ_i (b, q^k_b)$. Then, $(a, q^k_a) = (a, s^{\mu}_{a} + 1) \succ_i (b, q^k_b)$, and hence, i prefers lawyer a to b at step k, i.e., $aP^k_i b$. Since a's capacity is reduced by one at the beginning of step $k + 1$, a has at least one free slot at the end of step k under ψ. This contradicts that ψ is non-wasteful. Therefore, μ is mobility-free.

Let ψ be efficient in addition to being non-wasteful and individually rational. Then, μ is constrained efficient, and thus μ is also efficient by Lemma 1. ■

Proof of Proposition 4 Fix a profile $\succ$. Let ψ be a non-wasteful, individually rational, and group strategy-proof fixed-capacity mechanism, and let $K \subseteq N$ be the set of

naive defendants. Suppose for a contradiction that $\Gamma^{CC^\psi}(K, \succ)$ has a non-mobility-free equilibrium outcome, i.e., there is a profile of reported preferences $\bar{\succ}$, $i \in N$, and a lawyer a such that $(a, r+1) \succ_i \mu_i$ where $\mu = CC^\psi(\bar{\succ})$ and $r = s_a^\mu$. Note that by Theorem 3, $CC^\psi(\bar{\succ})$ is mobility-free under the reported profile $\bar{\succ}$. Then, because every naive defendant reports truthfully, any defendant who is mobile at $\mu = CC^\psi(\bar{\succ})$ under the true profile $\succ$ must be sophisticated. Consider the alternative report $\hat{\succ}_i$ of agent i where the top $r+1$ alternatives are $(a,1),(a,2),...,(a,r+1)$, while $\hat{\succ}_i$ and $\bar{\succ}_i$ coincide when defined on the rest of the alternatives. We will show that i is guaranteed a slot at lawyer a with a maximum congestion level of $r+1$ under $\hat{\succ} = (\hat{\succ}_i, \bar{\succ}_{-i})$ at CC^ψ, contradicting that $\bar{\succ}$ is a Nash equilibrium profile.

For any k, denote the fixed-capacity problems induced by $CC^\psi(\bar{\succ})$ and $CC^\psi(\hat{\succ})$ as $\overline{FCP}(k) = [\bar{\omega}^k, \bar{q}^k, \bar{P}^k]$ and $\widehat{FCP}(k) = [\hat{\omega}^k, \hat{q}^k, \hat{P}^k]$, respectively. Let $\bar{k}$ and $\tilde{k}$, with $\bar{k} \leq \tilde{k}$, be the first and last steps, respectively, such that the capacity of a at the congestion-cutting mechanism ψ under $\bar{\succ}$ is equal to $r+1$. Recall that the fixed-capacity problem at a given step of CC^ψ is recursively defined by the outcome of the preceding step, in terms of endowments and lawyer capacities, which in turn determines the stage preferences P_j. Since a has a capacity of at least $r+2$ for all steps $k < \bar{k}$ under $CC^\psi(\bar{\succ})$ and the preference profiles $\hat{\succ}$ and $\bar{\succ}$ agree on the set of all pairs that either involve lawyers other than a, or involve a only with $(a, r+2), (a, r+3), \ldots$, it follows that $\overline{FCP}(k) = \widehat{FCP}(k)$ for all such k as $CC^\psi(\bar{\succ})$ and $CC^\psi(\hat{\succ})$ run recursively in the same way until (excluding) step $\bar{k}$. Therefore, $\bar{\omega}^{\bar{k}} = \hat{\omega}^{\bar{k}}$ and $\bar{q}^{\bar{k}} = \hat{q}^{\bar{k}}$ as well.

$\overline{FCP}(\bar{k})$ and $\widehat{FCP}(\bar{k})$ differ only by agent i's induced preferences over lawyers $\bar{P}_i^{\bar{k}}$ and $\hat{P}_i^{\bar{k}}$. Moreover, by construction of $\hat{\succ}$, $\hat{P}_i^{\bar{k}}$ is derived from $\bar{P}_i^{\bar{k}}$ by moving lawyer a to the top of the ordering and keeping the relative ranking of all other lawyers the same.

Note that if i gets a slot in a at step $\bar{k}$ or any subsequent step under $CC^\psi(\hat{\succ})$, then, since ψ is individually rational and i cannot be worse off with respect to $\hat{\succ}_i$ as the algorithm proceeds, i is guaranteed lawyer a with a congestion level no more than $r+1$, finishing the proof.

Now, suppose for a contradiction that i never obtains a at step $\bar{k}$ or later under $CC^\psi(\hat{\succ})$. We first consider step $\bar{k}$. Let i be assigned to $b \neq a$ at $\widehat{FCP}(\bar{k})$ under $CC^\psi(\hat{\succ})$. We will first show that the tentative assignments at the end of step $\bar{k}$ under two preference profiles $\bar{\succ}$ and $\hat{\succ}$ should be the same by group strategy-proofness of ψ, i.e., the following holds: (1) i is assigned to b at $\overline{FCP}(\bar{k})$ under $CC^\psi(\bar{\succ})$, and (2) the tentative assignments of all agents at the end of step $\bar{k}$ are the same under $CC^\psi(\bar{\succ})$ and $CC^\psi(\hat{\succ})$.

To see that (1) holds, suppose that i is assigned to $l \neq b$ at $\overline{FCP}(\bar{k})$ under $CC^{\psi}(\succ)$ and recall that the difference between $\overline{FCP}(\bar{k})$ and $\widehat{FCP}(\bar{k})$ is only i's preferences. If $l\, \hat{P}_i^{\bar{k}}\, b$ (possibly $l = a$), i could manipulate ψ at the problem $\widehat{FCP}(\bar{k})$ by announcing $\bar{P}_i^{\bar{k}}$, which would contradict that ψ is strategy-proof. If $b\, \hat{P}_i^{\bar{k}}\, l$, (then $l \neq a$) since $\hat{P}_i^{\bar{k}}$ and $\bar{P}_i^{\bar{k}}$ coincide on the comparison of b and l (and hence, $b\, \bar{P}_i^{\bar{k}}\, l$), i could manipulate ψ at the problem $\overline{FCP}(\bar{k})$ by announcing $\hat{P}_i^{\bar{k}}$, which would contradict that ψ is strategy-proof. Having established (1), that i obtains the same lawyer under both problems, $\overline{FCP}(\bar{k})$ and $\widehat{FCP}(\bar{k})$, where only i's preference is different, the lawyer assignment of all agents should be the same under both problems.[46] Otherwise, at least one agent $j \neq i$ is better off when we move from one of the fixed-capacity problems to the other, giving an opportunity to i and j to jointly manipulate ψ by changing only i's preferences. Therefore, (2) holds.

If $\bar{k} = \tilde{k}$, (2) implies that a has a free slot at the end of step $\bar{k}$ at $\widehat{FCP}(\bar{k})$. However, this contradicts that ψ is non-wasteful since a is the top alternative in $\hat{P}_i^{\bar{k}}$ and i does not receive a at $\widehat{FCP}(\bar{k})$.

Consider the case $\bar{k} < \tilde{k}$. We already established that $\overline{FCP}(k) = \widehat{FCP}(k)$ for all $k \leq \bar{k}$ and $CC^{\psi}(\succ)$ and $CC^{\psi}(\hat{\succ})$ run recursively in the same way until (excluding) step $\bar{k}+1$. Therefore, $\bar{\omega}^{\bar{k}+1} = \hat{\omega}^{\bar{k}+1}$ and $\bar{q}^{\bar{k}+1} = \hat{q}^{\bar{k}+1}$ as well, and hence $\bar{q}_a^{\bar{k}+1} = \hat{q}_a^{\bar{k}+1} = r+1$. Moreover, $\overline{FCP}(\bar{k}+1)$ and $\widehat{FCP}(\bar{k}+1)$ differ only by i's preferences $\bar{P}_i^{\bar{k}+1}$ and $\hat{P}_i^{\bar{k}+1}$, where $\hat{P}_i^{\bar{k}+1}$ is derived from $\bar{P}_i^{\bar{k}+1}$ by moving lawyer a to the top of the ordering and keeping the relative ranking of all other lawyers the same. By using the same arguments for step $\bar{k}$, we can immediately tell that $\psi(\overline{FCP}(\bar{k}+1)) = \psi(\widehat{FCP}(\bar{k}+1))$ and $\psi_i(\widehat{FCP}(\bar{k}+1)) \neq a$. This contradicts that ψ is non-wasteful if $\bar{k}+1 = \tilde{k}$, since then a has a free slot at the end of step $\bar{k}+1$ at $\widehat{FCP}(\bar{k}+1)$ and a is the top alternative in $\hat{P}_i^{\bar{k}+1}$. If $\bar{k}+1 < \tilde{k}$, using the same arguments recursively, we can prove that $CC^{\psi}(\succ)$ and $CC^{\psi}(\hat{\succ})$ run in the same way until (including) step $\tilde{k}$ and $\psi_i(\widehat{FCP}(\tilde{k})) \neq a$. This contradicts that ψ is non-wasteful. ■

B Examples

Example 6. ***(The bound on manipulation is not tight).*** *Let $L = \{a, b, c\}$ and $n = 5$ and consider the profile $\succ$ discussed in Example 3. Define the mechanism φ as follows: φ lexicographically maximizes the number of clients assigned to lawyers a, b, and c,*

[46] This is true by the standard non-bossiness property known to follow from group strategy-proofness in a fixed-capacity setting.

respectively, in an efficient and mobility-free assignment. In case there are multiple assignments achieving that, φ chooses the one that lexicographically maximizes the welfare of defendants 1, 2, 3, and 4, respectively. Note that $\varphi(\succ) = \mu$ where $\mu^{-1}(a) = \{1, 2, 3\}$, $\mu^{-1}(b) = \{4, 5\}$, $\mu^{-1}(c) = \emptyset$. Now, consider $\bar{\succ}_1 : (a, 1), (a, 2), (b, 1), (b, 2), (b, 3)$, and the profile $\bar{\succ} = (\bar{\succ}_1, \succ_{-1})$. Clearly, $\varphi(\bar{\succ}) = \bar{\mu}$ where $(\bar{\mu})^{-1}(a) = \{1, 3\}$, $\bar{\mu}^{-1}(b) = \{2, 4\}$, $\bar{\mu}^{-1}(c) = \{5\}$. $\bar{\succ}$ is a successful manipulation under the true preference profile $\succ$, and $(a, 2)$ is the best assignment 1 could get by manipulating φ since it is always an option to assign agents 2 and 3 to a in case 1's preference does not allow her to share a with someone else in any mobility-free assignment. So, 1 does not have any strategy to achieve her best mobility-free assignment $(a, 1)$ under $\succ$. However, it is important to note that 2 or 3 can achieve their best mobility-free assignment $(b, 2)$ by just announcing that they prefer to have lawyer b regardless of his congestion level. ♦

Example 7. ***(An efficient mobility-free mechanism with a non-mobility-free equilibrium).*** *$n = 6$, $m = 7$. Only agent 1 is sophisticated and the true profile is as follows:*

$$(a, 1) \succ_1 (b, 1) \succ_1 (b, 2) \cdots$$
$$(c, 1) \succ_2 (a, 1) \succ_2 (a, 2) \cdots$$
$$(d, 1) \succ_3 (c, 1) \succ_3 (c, 2) \cdots$$
$$(e, 1) \succ_4 (c, 1) \succ_4 (c, 2) \cdots$$
$$(d, 1) \succ_5 (f, 1) \succ_5 (f, 2) \cdots$$
$$(e, 1) \succ_6 (g, 1) \succ_6 (g, 2) \cdots$$

Let $\varphi(\succ) = (b, a, c, c, d, e)$ be the mobility-free assignment at the true profile. If $\succ'_1$ is such that 1 top ranks $(b, 1)$, we have $\varphi(\succ') = (b, c, d, e, f, g)$, where $\succ' = (\succ'_1, \succ_{-1})$. Note that $\varphi(\succ')$ is not mobility-free under the original profile $\succ$ since 1 prefers $(a, 1)$ to $\varphi_1(\succ') = (b, 1)$ and no other agent is assigned to lawyer a.

To construct a mobility-free mechanism φ such that $\succ'$ is a Nash equilibrium of the game $\Gamma^{\varphi}(K, \succ)$, it suffices to ensure that for any strategy $\bar{\succ}_1$ against $\succ_{-1}$, $\varphi_1(\bar{\succ}) \neq (a, 1)$ where $\bar{\succ} = (\bar{\succ}_1, \succ_{-1})$.

For all $l \in \{f, g\}$, if defendant 1 top ranks $(l, 1)$ or $(a, 1)\bar{\succ}_1(l, 1)$,—that is, either $(l, 1)$ is the best alternative or $(a, 1)$ is the best and $(l, 1)$ is the second best—the final outcome of φ is (l, a, c, c, d, e). If she top ranks $(a, 1)\bar{\succ}_1(b, 1)$, the outcome is (b, a, c, c, d, e). If she top ranks $(d, 1)$ or $(a, 1)\bar{\succ}_1(d, 1)$, the outcome is (d, a, c, c, f, e). If she top ranks $(e, 1)$ or $(a, 1)\bar{\succ}_1(e, 1)$, the outcome is (e, a, c, c, d, g). If she top ranks $(c, 1)$, the outcome

is (c, a, d, e, f, g)*. If she top ranks* $(a, 1)\bar{\succ}_1(c, 1)\bar{\succ}_1(c, 2)$*, the outcome is* (c, a, c, e, d, g)*. If she top ranks* $(a, 1)\bar{\succ}_1(a, 2)$ *or* $(a, 1)\bar{\succ}_1(c, 1)\bar{\succ}_1(a, 2)$*, the outcome is* (a, a, c, c, d, e)*. Finally, if she top ranks* $(a, 1)\bar{\succ}_1(c, 1)\bar{\succ}_1(l, 1)$ *for* $l \in \{b, d, g, e, f\}$*, the outcome is the same as in the case where she top ranks* $(l, 1)$*. It is straightforward to check that for each case we covered above, the outcome is mobility-free and efficient under the corresponding profile* $\bar{\succ}$*. Since we covered all the strategies of agent 1 and she is never the only client of lawyer* a*,* $\succ'$ *is a Nash equilibrium.* ♦

Example 8. ***(An efficient mobility-free mechanism with no equilibrium).*** $n = 3$*,* $m = 3$*. Only agents 1 and 2 are sophisticated and the true profile is as follows:*

$$(c, 1) \succ_1 (b, 1) \succ_1 (b, 2) \cdots$$
$$(c, 1) \succ_2 (b, 1) \succ_2 (a, 1) \cdots$$
$$(c, 1) \succ_3 (b, 1) \succ_3 (c, 2) \cdots$$

We consider the following mechanism. For any announced profile $\bar{\succ} = (\bar{\succ}_1, \bar{\succ}_2, \succ_3)$*, let* $\varphi^{lex}(\bar{\succ})$ *be the lexicographically smallest assignment (with* $a < b < c$*) among the assignments that are both efficient and mobility-free under the announced profile, i.e., if* (a, a, a) *is efficient and mobility-free under* $\bar{\succ}$*,* $\varphi^{lex}(\bar{\succ}) = (a, a, a)$*, if not and* (a, a, b) *is efficient and mobility-free under* $\bar{\succ}$*,* $\varphi^{lex}(\bar{\succ}) = (a, a, b)$*, and so on. This mechanism is well-defined as every profile admits at least one efficient and mobility-free assignment.*

Note that there are 27 possible assignments. Since φ^{lex} *is mobility-free, for any announced profile* $\bar{\succ}$*, each defendant will get one of the top-*3 *assignments in their announced preference. Therefore, the naive defendant (3) will get* $(c, 1)$*,* $(c, 2)$*, or* $(b, 1)$*. This renders any assignment at which 3 gets* a*, or shares* b *with at least one other defendant, or shares* c *with the other two defendants infeasible as an outcome under* $\bar{\succ}$*. Moreover, defendant 3 is mobile at the assignments* (a, a, b)*,* (a, c, c)*, and* (c, a, c) *under* $\bar{\succ}$*. Hence,* $\varphi^{lex}(\bar{\succ})$ *is one of the following assignments:*

$$\{(a, a, c), (a, b, c), (a, c, b), (b, a, c), (b, b, c), (b, c, c), (c, a, b), (c, b, c), (c, c, b)\}.$$

Now, suppose $\bar{\succ} = (\bar{\succ}_1, \bar{\succ}_2, \succ_3)$ *is a Nash equilibrium strategy profile. Recall that* φ^{lex} *is a mobility-free mechanism, and hence, each defendant is guaranteed to get one of her top-*3 *assignments with respect to her announced preference. Consequently, 1 guarantees* $(c, 1)$*,* $(b, 1)$*, or* $(b, 2)$ *if she announces* $\succ_1$ *and 2 guarantees* $(c, 1)$*,* $(b, 1)$*, or* $(a, 1)$ *if she announces* $\succ_2$*. Then, any assignment at which defendant 1 gets lawyer* a*,*

$(b,2)$, *or* $(c,2)$ *and defendant 2 gets* $(c,2)$ *cannot be the outcome of* $\varphi^{lex}(\bar{\succ})$. *Hence, eliminating 7 more assignments, we have* $\varphi^{lex}(\bar{\succ}) \in \{(b,a,c),(c,a,b)\}$.

Case 1: $\varphi^{lex}(\bar{\succ}) = (b,a,c)$. *Consider the preference* $\hat{\succ}_1 : (c,1)\hat{\succ}_1(c,2)\hat{\succ}_1(c,3)\cdots$ *and the profile* $\hat{\succ} = (\hat{\succ}_1, \bar{\succ}_2, \succ_3)$. *At any mobility-free assignment, defendant 1 gets lawyer c and defendant 3 gets lawyer b or c. Then,* (c,a,b) *is the lexicographically minimal assignment among feasible outcomes of* $\varphi^{lex}(\hat{\succ})$. *Moreover, since* (b,a,c) *is mobility-free and efficient under* $\bar{\succ}$ *and defendant 2 keeps her preference in* $\hat{\succ}$, (c,a,b) *is mobility-free and efficient under* $\hat{\succ}$. *Therefore,* $\varphi^{lex}(\hat{\succ}) = (c,a,b)$. *This contradicts that* $\bar{\succ}_1$ *is a best response to* $(\cdot, \bar{\succ}_2, \succ_3)$ *as we assumed* $\bar{\succ}$ *is a Nash equilibrium profile.*

Case 2: $\varphi^{lex}(\bar{\succ}) = (c,a,b)$. *Then* $(c,1)\bar{\succ}_1(a,2)$ *by mobility-freeness, and moreover,* $(c,1) \succ_1 (b,1)$ *by efficiency as otherwise* (b,a,c) *Pareto dominates* (c,a,b). *Suppose for a contradiction that* $(a,1)$ *is one of the top-2 alternatives in* $\bar{\succ}_1$. *Then, if 2 announces lawyer-c lexicographic preference instead of* $\bar{\succ}_2$ *as a response, the outcome of* φ^{lex} *will be* (a,c,b), *making 2 better off and contradicting that* $\bar{\succ}$ *is a Nash equilibrium.*

Therefore, $(c,1)$ *is the top alternative in* $\bar{\succ}_1$ *and either* $(c,2)$ *or* $(b,1)$ *is the second best alternative. In the former case, 2 can announce lawyer-b lexicographic preference as a response and guarantee* $(b,1)$ *since in that case the outcome of the mechanism will be* (c,b,c). *This contradicts that* $\bar{\succ}$ *is a Nash equilibrium profile as* $(b,1) \succ_2 (a,1)$. *In case* $(b,1)$ *is the second-best alternative in* $\bar{\succ}_1$, *then, if* (c,a,b) *is efficient and mobility-free,* (b,a,c) *is also efficient and mobility-free. This contradicts that* $\varphi^{lex}(\bar{\succ}) = (c,a,b)$.
♦

References

Atila Abdulkadiroğlu and Tayfun Sönmez. School choice: A mechanism design approach. *American Economic Review*, 93(3):729–747, 2003.

Atila Abdulkadiroğlu and Tayfun Sönmez. House allocation with existing tenants. *Journal of Economic Theory*, 88(2):233–260, 1999.

Amanda Agan, Matthew Freedman, and Emily Owens. Is your lawyer a lemon? incentives and selection in the public provision of criminal defense. *Review of Economics and Statistics*, 103(2):294–309, 2021.

Kemal Kıvanç Aköz, Emre Doğan, Onur Kesten, and Danisz Okulicz. Stability as right to counsel of choice: A lawyers' matching problem. *Games and Economic Behavior*, 152:1–22, 2025.

James M Anderson and Paul Heaton. How much difference does the lawyer make: The effect of defense counsel on murder case outcomes. *Yale Law Journal*, 122:154, 2012.

Orhan Aygün and Günnur Ege Bilgin. Placement with assignment guarantees and semi-flexible capacities. Technical report, University of Bonn and University of Mannheim, Germany, 2023.

Sean Baker. The saving grace of public defense? is the "client-choice" method a cure-all for problems that plague this overburdened system? 50 j. marshall l. rev. 307 (2017). *UIC Law Review*, 50(2):6, 2017.

Anna Bogomolnaia and Hervé Moulin. Fair congested assignment. *Mathematics of Operations Research*, 2025. doi: 10.1287/moor.2024.0581. Published online: 7 May 2025.

Kemble Rhea Brecher. Sixth amendment and the right to counsel. *University of Pennsylvania Law Review*, 136:1957, 1987.

Emily D. Buehler. Justice expenditure and employment in the united states, 2017. *Bureu of Justice Statistics*, 2021.

Jiehua Chen, Jiong Guo, and Yinghui Wen. Algorithms and complexity for congested assignments. *arXiv preprint arXiv:2312.07431*, 2023.

Andrew Copland. School choice and class size externalities. *Available at SSRN 4346860*, 2023.

David Gale and Lloyd S Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

Alfredo Garcia. The right to counsel under siege: Requiem for an endangered right. *American Criminal Law Review*, 29:35, 1991.

Brandon L Garrett. Judging innocence. *Columbia Law Review*, pages 55–142, 2008.

John Haltiwanger and Michael Waldman. Rational expectations and the limits of rationality: An analysis of heterogeneity. *American Economic Review*, 75(3):326–340, 1985.

Ernan Haruvy, Alvin E Roth, and M Utku Ünver. The dynamics of law clerk matching: An experimental and computational investigation of proposals for reform of the market. *Journal of Economic Dynamics and Control*, 30(3):457–486, 2006.

Michael Heise, Nancy J King, and Nicole A Heise. State criminal appeals revealed. *Vanderbilt Law Review*, 70:1939, 2017.

Erinn Herberman and Tracey Kyckelhahn. State government indigent defense expenditures, fy2008–2012 - updated. *Bureu of Justice Statistics*, 2014.

Elisabetta Iossa and Bruno Jullien. The market for lawyers and quality layers in legal services. *The RAND Journal of Economics*, 43(4):677–704, 2012.

Radha Iyengar. An analysis of the performance of federal indigent defense counsel. Working Paper 13187, National Bureau of Economic Research, June 2007. URL `http://www.nber.org/papers/w13187`.

Behrang Kamali Shahdadi. Matching with moral hazard: Assigning attorneys to poor defendants. *American Economic Journal: Microeconomics*, 10(3):1–33, 2018.

Hideo Konishi, Michel Le Breton, and Shlomo Weber. Equilibria in a model with partial rivalry. *Journal of Economic Theory*, 72(1):225–237, 1997.

Lyn Langton and Donald J. Farole. Public defender offices, 2007. *Bureu of Justice Statistics Selected Findings*, 2009.

Lyn Langton and Donald J. Farole. County-based and local public defender offices, 2007. *Bureu of Justice Statistics Special Report*, 2010.

Norman Lefstein. Securing reasonable caseloads. *Ethics and Law in Public Defense*, 2011.

Igal Milchtaich. Congestion games with player-specific payoff functions. *Games and Economic Behavior*, 13(1):111–124, 1996.

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

A New York State Office of Indigent Legal Services. Estimate of the cost of compliance with maximum national caseload limits in upstate new york - 2013 update. *Technical report*, 2014.

M. Elain Nugent-Borakove and Franklin Cruz. The power of choice. the implication of a system where indigent defendants choose their own counsel. *The Justice Management Institute Report*, 2017.

Charles J Ogletree Jr. Beyond justifications: Seeking motivations to sustain public defenders. *Harvard Law Review*, 106:1239, 1992.

Parag A Pathak and Tayfun Sönmez. Leveling the playing field: Sincere and sophisticated players in the boston mechanism. *American Economic Review*, 98(4):1636–1652, 2008.

William Phan, Ryan Tierney, and Yu Zhou. Crowding in school choice. *American Economic Review*, 114(8):2526–2552, 2024.

Doug Pyper, Georgina Sturge, Sally Lipscomb, and Sue Holland. Spending of the mininstry of justice on legal aid. *Debate Pack*, 2020.

Szilvia Pápai. Strategyproof assignment by hierarchical exchange. *Econometrica*, 68 (6):1403–1433, 2000.

Thomas Quint and Martin Shubik. A model of migration. Cowles Foundation Discussion Paper 1088, Cowles Foundation for Research in Economics, Yale University, 1994. URL `https://cowles.yale.edu/sites/default/files/2022-08/d1088.pdf`.

John Rappaport. The structural function of the sixth amendment right to counsel of choice. *The Supreme Court Review*, 2016(1):117–156, 2017.

Robert W Rosenthal. A class of games possessing pure-strategy nash equilibria. *International Journal of Game Theory*, 2(1):65–67, 1973.

Alvin E Roth and Elliott Peranson. The redesign of the matching market for american physicians: Some engineering aspects of economic design. *American Economic Review*, 89(4):748–780, 1999.

Alvin E Roth, Tayfun Sönmez, and M Utku Ünver. Kidney exchange. *The Quarterly Journal of Economics*, 119(2):457–488, 2004.

Stephen J Schulhofer and David D Friedman. Rethinking indigent defense: promoting effective representation through consumer sovereignty and freedom of choice for all criminal defendants. *American Criminal Law Review*, 31:73, 1993.

Lloyd Shapley and Herbert Scarf. On cores and indivisibility. *Journal of Mathematical Economics*, 1(1):23–37, 1974.

Yotam Shem-Tov. Make or buy? the provision of indigent defense services in the united states. *The Review of Economics and Statistics*, 104(4):819–827, 2022.

Stephen J Spurr. How the market solves an assignment problem: The matching of lawyers with legal claims. *Journal of Labor Economics*, 5(4, Part 1):502–532, 1987.

Tayfun Sönmez and M. Utku Ünver. House allocation with existing tenants: A characterization. *Games and Economic Behavior*, 69(2):425–445, 2010.

Peter W Tague. An indigent's right to the attorney of his choice. *Stanford Law Review*, 27:73, 1974.

Ryan Tierney. The problem of multiple commons: A market design approach. *Games and Economic Behavior*, 114:1–27, 2019.

Peter Troyan and Thayer Morrill. Obvious manipulations. *Journal of Economic Theory*, 185:104970, 2020.

Bill Whitehurst. Gideon's broken promise: America's continuing quest for equal justice. *A Report on the American Bar Association's Hearings on the Right to Counsel in Criminal Proceedings*, 2004.