Fragile Stable Matchings

Kirill Rudov*

August 5, 2024

Abstract

We show how fragile stable matchings are in a decentralized one-to-one matching setting. The classical work of Roth and Vande Vate (1990) suggests simple decentralized dynamics in which randomly-chosen blocking pairs match successively. Such decentralized interactions guarantee convergence to *a* stable matching. Our first theorem shows that, under mild conditions, *any* unstable matching—including a small perturbation of a stable matching—can culminate in *any* stable matching through these dynamics. Our second theorem highlights another aspect of fragility: stabilization may take a long time. Even in markets with a unique stable matching, where the dynamics always converge to the same matching, decentralized interactions can require an exponentially long duration to converge. A small perturbation of a stable matching may lead the market away from stability and involve a sizable proportion of mismatched participants for extended periods. Our results hold for a broad class of dynamics.

Keywords: Fragility, Stability, Decentralized Matching, Market Design.

*Department of Economics, UC Berkeley (email: krudov@berkeley.edu). I am grateful to Leeat Yariv for her guidance in developing this project. I would like to thank Noga Alon, Ata Atay, Arjada Bardhi, Péter Biró, Sylvain Chassang, Federico Echenique, Andrew Ferdowsian, Faruk Gul, Navin Kartik, Alessandro Lizzeri, Sofia Moroni, Xiaosheng Mu, Pietro Ortoleva, Wolfgang Pesendorfer, Doron Ravid, Joseph Root, Alvin Roth, Eran Shmaya, Alex Teytelboym, Can Urgun, Vijay Vazirani, and various seminar and conference participants for helpful comments and insightful discussions.

1. Introduction

1.1. Overview. We examine the fragility of stable matchings in a decentralized one-to-one matching environment.[1] Most of the existing literature studies centralized markets: the medical residency match, school allocations, and others. However, many markets are not fully centralized: marriage markets, the market matching junior economists and academic positions, college admissions, and so on. Furthermore, decentralized interactions often precede or follow centralized markets. Even when a market relies on a stable centralized clearinghouse, ex-post preference shocks, changes in market composition, or small implementation errors may lead to an unstable matching, close to or distant from the intended stable matching. How fragile are stable matchings with respect to small perturbations? In general, when a matching in place is unstable, what is the set of outcomes decentralized interactions can generate? Even if convergence to a desired stable matching is guaranteed, what are efficiency costs of decentralized interactions, in terms of duration of instability and the proportion of market participants who spend long periods mismatched? These questions are at the heart of this paper.

We consider a decentralized process in which randomly-chosen blocking pairs successively match with each other, potentially breaking their former matches. These dynamics always culminate in a stable matching and impose low sophistication requirements on market participants. Theorem 1 shows that, under mild conditions, starting from *any* unstable matching, these dynamics attain *any* stable matching with positive probability. In particular, even small perturbations of stable matchings may lead to *any* stable matching. Thus, stable matchings are fragile. Simulations suggest that, in fact, even a minimal deviation from a stable matching might converge with substantial probability to another, possibly vastly different stable matching. Another insight of this paper is that the stabilization dynamics may involve many mismatched participants for long durations. Theorem 2 shows that, even in markets with a unique stable matching, where the dynamics always converge to the same stable matching, stabilization may take an exponentially long time. This result highlights another fragility aspect: a small perturbation of a stable outcome may lead the market away from stability for a long stretch of time. Using simulations, we show that, whether or not the stable matching is unique, the stabilization process typically strays from stability, takes a long time to regain it, and involves many mismatched participants for substantial periods.

[1] A matching is stable if there is no pair of participants that prefer each other to their partners in the matching; such a pair is called a blocking pair (see Roth and Sotomayor, 1992 for details).

Altogether, our results imply the fragility of stable matchings. Our analysis illustrates the importance of taking into account decentralized interactions that precede or follow centralized markets and casts doubt on empirical studies of decentralized markets that estimate parameters in one snapshot of time, relying on stability as an identification assumption.

Stability is a central concept in matching theory. It is defined by the absence of blocking pairs, which are pairs of participants who prefer each other over their current partners (Gale and Shapley, 1962). This definition is motivated by the concern that some participants may profitably deviate and circumvent the intended matching by forming blocking pairs. Such a concern was part of the impetus for introducing stable centralized clearinghouses.

It is commonly believed that if an unstable outcome occurs, decentralized interactions eventually yield stability. Going back to Knuth (1976), the literature has sought to provide a theoretical framework for understanding how stable outcomes emerge. The classical work of Roth and Vande Vate (1990) suggests simple and natural decentralized dynamics generating stability. They show that in one-to-one markets, for any unstable matching, there exists a finite sequence of blocking pairs that generates a stable matching. Consequently, when blocking pairs are formed at random in each step, stability is guaranteed. A benefit of these dynamics is that convergence to stability is possible even when market participants have limited sophistication and possess only local information—participants need to identify their own blocking pairs, but not much more.

We consider a broad class of such dynamics, allowing for non-uniform and time-dependent formation of blocking pairs, for which convergence to stability is guaranteed (cf. Roth and Vande Vate, 1990). This class includes dynamics where, at each step, a participant randomly chosen from those with at least one blocking partner selects her *best* blocking partner. The probability that a particular blocking pair of market participants matches can be quite general, reflecting factors such as the likelihood that these participants would meet, or their incentives to form a match, potentially driven by cardinal payoffs.

We study the robustness of stable matchings with respect to arbitrary, possibly minimal, perturbations. Such perturbations may arise in *both decentralized and centralized* settings. Over time, participants' preferences may shift, market composition may change, or matched participants might end their partnerships. For instance, in labor markets, a family shock may generate new geographical preferences for workers; new positions may appear and workers may either become available or retire; an employer and an employee might mutually agree to terminate the employment relationship. While the set of stable matchings may also be altered by such changes, they could all lead to instability. In particular, some participants that should be paired under *current* market conditions might be mismatched or unmatched.

We demonstrate how fragile stable matchings are under conditions conducive to their robustness. Specifically, we allow for an initial deviation from stability, stemming from an arbitrary shock, to be small. In addition, we suppose that there are no further shocks that might impede the stabilization process. Even under these favorable conditions, our findings reveal that stable matchings are fragile.

In some situations, decentralized interactions cannot break certain partnerships. Imagine a two-sided market divided into two submarkets, New York and Los Angeles, in which every Angeleno prefers any Angeleno to any New Yorker, and vice versa. Suppose that the New York submarket has a unique stable matching and the Los Angeles submarket has two stable matchings. The entire market naturally has two stable matchings as well. If all Angelenos are paired in accordance with one of the two stable matchings for the Los Angeles submarket, and therefore for the entire market, decentralized interactions cannot unmatch any such pair of matched Angelenos. As a result, the stabilization dynamics cannot attain another stable matching in the entire market. Here, Angelenos can be matched in a stable way inside the group so that every insider prefers her stable partner to anyone outside Los Angeles. We call such a group a *fragment*. In this example, New Yorkers form a fragment as well. A fragment is *trivial* if all stable matchings in the entire market match participants inside the fragment in the same way. In particular, Angelenos constitute a *non-trivial* fragment, while New Yorkers form a trivial fragment.

If a market has a non-trivial fragment, some unstable matchings cannot yield certain stable matchings. Theorem 1 shows that the reverse is also true. When there are no non-trivial fragments, *any* unstable matching can yield *any* stable matching through decentralized interactions. Non-trivial fragments are the only restraints on the stabilization dynamics. The absence of non-trivial fragments is a mild condition: simulations suggest that they are rare in large random markets. Thus, in most markets, the decentralized process is fluid enough to attain any stable matching.

Theorem 1 suggests one type of fragility of stable matchings. A small perturbation of a stable matching may culminate in *any* stable matching, close to or distant from the original stable matching. The celebrated Deferred Acceptance mechanism of Gale and Shapley (1962), which is used to match doctors to hospitals and students to schools, aims at implementing an extremal stable matching, the best stable matching for one market side, doctors and students, respectively. Our results imply that any perturbation of such a clearinghouse's outcome need not revert back to the intended extremal stable matching; decentralized interactions can instead lead to the other extremal stable matching, or anything in between.

We employ simulations to investigate the stable matchings that have stronger drawing power in random markets with multiple stable matchings. In order to inspect the effects of minimal perturbations to stability, we initialize markets at an *almost stable* matching, which is one blocking pair away from stable. For presentation simplicity, suppose a blocking pair is chosen uniformly at random at each step of the dynamics.[2] The simulations show that almost stable matchings—even those corresponding to an extremal stable matching—converge with substantial probability to a stable matching distant from the slightly perturbed matching.

Fragility may take other forms: the stabilization dynamics may take a long time and involve a sizable fraction of participants mismatched for a long duration. To isolate these other forms of fragility, we focus on markets with a unique stable matching. For such markets, the dynamics always converge to the same matching, the unique stable one. In fact, there is work suggesting that large markets have small cores in certain settings.[3]

Theorem 2 shows that for a large class of markets with a unique stable matching, even a small deviation from stability entails an exponentially long stabilization process. Specifically, we start with *any* market with a unique stable matching. We show that the market can be augmented with a small number of market participants so that the resulting market still features a unique stable matching, but where for most matchings on any path to stability, there are many blocking pairs. Importantly, considerably more blocking pairs are "destabilizing," moving the augmented market away from stability, than "stabilizing," moving the augmented market towards stability. The stabilization dynamics can then be associated with a random walk that is heavily biased in the direction of instability. This bias substantially decelerates the convergence, making it exponential.

We illustrate that fragments accelerate convergence. For example, consider a labor market in which participants on each market side have assortative preferences—say, firms have the same ranking of workers based on their college GPA and workers rely on the same company

[2]Analogous results hold for other dynamics, specifically for dynamics where, at each step, a randomly-chosen participant forms a match with her best blocking partner.

[3]Several empirical papers document small cores for *reported* preferences in some markets (e.g., Roth and Peranson, 1999). There are also theoretical studies identifying conditions under which large markets have an essentially unique stable matching (see Immorlica and Mahdian, 2005, Kojima and Pathak, 2009, and Ashlagi, Kanoria, and Leshno, 2017). Of course, our first theorem has bite as long as there are multiple stable matchings, even when they are not numerous. In addition, there is growing evidence that reported preferences might differ significantly from truthful ones; see Artemov, Che, and He (2023), Echenique et al. (2022), and references therein. Furthermore, recent papers suggest that cores are large in certain settings, including large imbalanced markets (Biro et al., 2022; Hoffman, Levy, and Mossel, 2023; Rheingans-Yoo, 2024; Brilliantova and Hosseini, 2022), with possibly correlated preferences.

ratings site to rank firms. Then, the best firm and the best worker form a trivial fragment. In addition, the top two firms and the top two workers constitute a trivial fragment that nests the previous one, and so forth. This market belongs to a general class of markets having a nested structure of fragments. We show that for any market in this class, the stabilization dynamics take polynomial time. However, in view of our second theorem, even such markets are fragile when augmented with a small fraction of new participants.

We use simulations to inspect convergence speeds in random markets with and without a unique stable matching. The simulations suggest that, regardless of whether stable matchings are unique, and even when starting from almost stable matchings, the stabilization dynamics typically move far away from stability. Consequently, stabilization takes a very long time, and involves long durations with a sizable fraction of market participants being mismatched.

Taken together, our results suggest the importance of accounting for decentralized interactions, even in the presence of stable matching clearinghouses. Even in an idealized setting, free from implementation concerns, shifts in participants' preferences or other previously discussed changes in market conditions can naturally occur after the desired stable matching is implemented. While these changes may result in a minor deviation from the targeted stable outcome initially, decentralized interactions may further destabilize the market significantly. Such undesirable dynamics may affect a large fraction of market participants for a long period of time, potentially failing to attain the intended stable matching.

From a market design perspective, interventions directed at reducing efficiency costs of decentralized interactions might prove useful. These interventions may include designing specific rules governing how participants interact with each other or introducing restrictions that reduce market activity. Nonetheless, such interventions inevitably come at the expense of flexibility for decentralized interactions to adjust to market conditions. As market conditions evolve, there may be many unsatisfied participants unable to improve their situations only due to these imposed restrictions. This highlights a trade-off, an analysis of which may be valuable for future research and market design: while we may not want decentralized interactions to be too flexible, we may still need them to be flexible enough to adapt to changing market conditions.[4]

[4]As suggestive evidence of impacts of decentralized interactions even in *centralized* markets, there is a growing concern regarding the highly and increasingly mobile U.S. doctor workforce, especially for early-stage doctors. For instance, more than 40% of practice positions turned over for radiology between 2014 and 2018 (Santavicca et al., 2021), with increasing annual turnover rates. In comparable periods, more than half of practice positions turned over for dermatology (Cwalina et al., 2022), and over one-third for both otolaryngology (Sheth et al., 2023) and ophthalmology (Patel et al., 2023). For trainees, transfers between residency programs are expectedly less common due to strict restrictions, such as requiring approval from a

An alternative, and possibly complementary, approach would be to devise a dynamic reassignment clearinghouse that addresses instabilities on a regular basis. As noted earlier, a one-time centralized assignment is no panacea, as it does not account for post-matching decentralized interactions. A dynamic centralized mechanism could preemptively resolve certain instabilities before they become significant, whether arising from decentralized interactions or changing market conditions. The existing literature rarely considers dynamic matching mechanisms, and their consideration may be beneficial.

Our results also have implications for empirical work on matching. A growing empirical literature uses stability of observed matches as an identification assumption to estimate preferences in decentralized markets (see the surveys by Fox, 2009 and Chiappori and Salanié, 2016, as well as Boyd et al., 2013). Our findings imply that observed matchings might be far from stable for a long duration. This may introduce biases into estimates, of unclear directions and magnitudes, raising caution for the interpretation of those estimates. Highly mobile markets, where significant turnover may serve as an indicator of underlying instabilities, are especially vulnerable to this concern; for instance, these include the market for public school teachers analyzed in Boyd et al. (2013).[5] Although stability is an appealing structural assumption, empirical researchers may need to exercise additional rigor in verifying its validity in practical applications, potentially by inspecting historical turnover rates, conducting surveys, or examining post-matching behavior.

We provide additional implications in the discussion section at the end of the paper.

1.2. Related Literature. To the best of our knowledge, no prior work has examined the fragility of stable matchings with respect to arbitrary perturbations through the lens of decentralized interactions. In a somewhat related study, Jackson and Watts (2002) use similar dynamics to investigate the evolution of networks in the presence of mutations that might randomly add or delete links. Applied to matching problems, their results show that

current program director. However, transfers are an observed phenomenon, especially in psychiatry, general surgery, internal medicine, and family medicine (Accreditation Council for Graduate Medical Education, n.d.). For instance, in psychiatry, more than 5% of residents have been reported to transfer annually over recent years, with a total annual rate of 7-8% leaving their program before completion (Wang et al., 2022). While appealing for limiting mobility, the imposed restrictions may contribute to lower satisfaction among trainees and translate to higher attrition rates during residency and later in careers, ultimately imposing significant costs on the healthcare system.

[5]The issue of teacher turnover has gained increasing attention in the U.S. Over the past decade, around 8% of public school teachers have transferred to different schools each year, with an additional similar percentage leaving the teaching profession annually (Taie and Lewis, 2023). For further evidence of instability in markets for teachers in the U.S. and other countries, see Shure and Weingarten (2024) and references therein.

the support of the limiting stationary distribution coincides with the set of stable matchings; in this sense, no stable matching is more "fragile" than others in the face of such mutations.[6] Close to our work, the recent and growing computer science literature studies robust solutions to stable matchings in the presence of incomplete information about preferences (Aziz et al., 2020), stable matchings that are most tolerant to special classes of errors (Genc et al., 2019; Gangam et al., 2022), and algorithms to find a stable matching—after certain changes—that is as close as possible to an original stable matching (see Boehmer, Heeger, and Niedermeier, 2022 and references therein), to name a few. More broadly, the current paper relates to the fragility of complex economic systems (e.g., see Elliott and Golub, 2022).[7]

A few papers model a decentralized matching process by which stable outcomes are created. Roth and Vande Vate (1990) propose simple dynamics of pairwise blocking.[8,9] Some recent studies impose additional structure on the matching process or markets to define a decentralized market game in which agents interact strategically; see Haeringer and Wooders (2011), Ferdowsian, Niederle, and Yariv (2023), and references therein. They seek to identify conditions under which stable outcomes arise as equilibrium outcomes. We contribute to this literature by shifting the focus to the robustness of stability. As a by-product, we characterize markets in which the Roth and Vande Vate (1990) dynamics can yield any stable outcome.[10]

Several experimental papers study frictionless decentralized markets with complete information (Echenique, Robinson-Cortés, and Yariv, 2023; Pais, Pintér, and Veszteg, 2020).[11] These studies examine convergence to stable matchings and their selection assuming that participants start with an empty matching in which no one is matched. Stable outcomes are prevalent in lab markets. In markets with multiple stable matchings, a non-extremal stable

[6] Newton and Sawa (2015) show that for alternative mutation models, some stable matchings might not be stochastically stable.

[7] Kojima (2011) examines a mechanism for the school choice setting that is robust to a combined manipulation, where a student first misreports his preferences and then blocks the implemented matching.

[8] Similar stabilization results were later obtained for roommate markets (Chung, 2000; Diamantoudi, Miyagawa, and Xue, 2004; Inarra, Larrea, and Molis, 2008), many-to-one markets with couples (Klaus and Klijn, 2007), many-to-many markets without contracts (Kojima and Ünver, 2008) and with contracts (Millán and Pepa Risma, 2018), markets with incomplete information (Lazarova and Dimitrov, 2017; Chen and Hu, 2020), and supply chain networks (Rudov, 2024).

[9] Another stream of the literature models market participants as farsighted, see Herings, Mauleon, and Vannetelbosch (2020) and references therein.

[10] The search and matching literature pursues a rather different approach to the emergence of stable matchings, in which frictions are modeled explicitly and play a crucial role (see the survey by Chade, Eeckhout, and Smith, 2017 and references therein). Both approaches are used in empirical work to estimate preferences in decentralized markets (Chiappori and Salanié, 2016; Chiappori, 2020).

[11] Pais, Pintér, and Veszteg (2020) also study the impact of commitment, search costs, and limited information. Agranov et al. (2023) show that incomplete information hinders stability in markets with transfers.

matching—which presents a compromise between the two sides of the market—emerges most frequently (Echenique, Robinson-Cortés, and Yariv, 2023). For robustness purposes, we consider any starting matching. We show that even a minimal deviation from an extremal stable matching may lead to a completely different stable matching with substantial probability.

Although the convergence dynamics and the selection of stable matchings have recently received attention in the theoretical literature, mostly in computer science, they remain largely understudied. Ackermann et al. (2011) construct a particular sequence of markets, with an increasing number of stable matchings, and the corresponding sequence of starting matchings to illustrate an exponential lower bound for the convergence time assuming uniform Roth and Vande Vate (1990)-type random dynamics. They also identify a class of markets for which the convergence is polynomial.[12] Motivated by fragility, we extend their techniques to prove that, for a large class of markets with a unique stable matching, small deviations from stability lead to the exponential stabilization under a wide range of dynamics. We consider a broader class of markets for which the stabilization is polynomial. Moreover, we examine how the stabilization dynamics affect market participants. Biró and Norman (2013) connect the dynamics to Markov chains and show how to calculate the absorption probabilities of various stable matchings.[13] We use this connection to quantify the fragility of stable matchings in smaller markets. More generally, the current paper suggests that fragments might be a key determinant of the dynamics.

Several studies design re-equilibration mechanisms for labor markets in which matchings are destabilized by retirements or new entries. Blum, Roth, and Rothblum (1997) propose a procedure, closely related to the deferred acceptance algorithm, that regains stability in case of such disruptions. Combe et al. (2022) introduce static mechanisms involving the assignment of new workers and reassignment of existing ones to rebalance markets that suffer from distributional issues—in their empirical application, the unequal distribution of experienced public school teachers across regions of France.[14] While their focus is on distributional problems, our results indicate that reassignment may also be important for addressing various instabilities arising from decentralized interactions. Since these issues are dynamic in nature, static centralized mechanisms may not suffice, requiring the design of dynamic reassignment clearinghouses.[15]

[12]Later, Hoffman, Moeller, and Paturi (2013) expanded this class by using a graph-theoretic approach.

[13]See also Boudreau (2011, 2012).

[14]See Combe, Tercieux, and Terrier (2022) and references in these two papers.

[15]Akbarpour, Li, and Gharan (2020) study dynamic mechanisms for networked markets in which agents enter or exit stochastically over time. See Baccara and Yariv (2023) for a review of the recent developments

2. The Model

2.1. Basic Definitions. A *matching market* is a triplet $\mathcal{M} = (F, W, \succ)$ composed of a finite set of firms $F = \{f_i\}_{i\in[n]}$, where $[n] = \{1, 2, \dots, n\}$, a finite set of workers $W = \{w_j\}_{j\in[m]}$, and a profile of strict preferences $\succ = \{\succ_{f_i}, \succ_{w_j}\}_{(i,j)\in[n]\times[m]}$.[16] That is, each firm f_i is endowed with a strict preference relation $\succ_{f_i}$ over workers and being single, i.e., $W \cup \{f_i\}$. Similarly, each worker w_j is endowed with a strict preference relation $\succ_{w_j}$ over $F \cup \{w_j\}$. Throughout, we focus on markets where all worker-firm pairs are mutually acceptable, i.e., $w_j \succ_{f_i} f_i$ and $f_i \succ_{w_j} w_j$ for all i, j.[17]

A *matching* is a one-to-one function $\mu : F \cup W \to F \cup W$ that (i) assigns to each firm f_i either a worker or herself, $\mu(f_i) \in W \cup \{f_i\}$; (ii) assigns to each worker w_j either a firm or himself, $\mu(w_j) \in F \cup \{w_j\}$; (iii) is of order two, i.e, $\mu(f_i) = w_j$ if and only if $\mu(w_j) = f_i$. For a given matching μ, a pair (f_i, w_j) is said to form a *blocking pair* if firm f_i and worker w_j are not matched to one another at μ, but prefer each other over their assigned partners, i.e., $w_j \succ_{f_i} \mu(f_i)$ and $f_i \succ_{w_j} \mu(w_j)$. A matching is *stable* if it has no blocking pairs.

2.2. Decentralized Process. A folk argument suggests that if an unstable outcome is realized, decentralized interactions eventually attain stability. The leading model of such a decentralized process is Roth and Vande Vate (1990). They show that for any unstable matching, there is a finite sequence of successive matches arising from blocking pairs that generates a stable matching. Hence, the decentralized process of allowing randomly-chosen blocking pairs to match yields stability with certainty. An advantage of this process is that convergence to stability is ensured even when agents have limited sophistication and know only local information about others' preferences.[18] Theoretically, these dynamics are employed to study the formation of networks (Jackson and Watts, 2002), to justify stability as a suitable identification assumption for estimating preferences in decentralized matching markets (Boyd et al., 2013), to provide the decentralized foundation of incomplete-information stability (Chen and Hu, 2020), and more.

in dynamic matching.

[16] For convenience, we label two sides as firms and workers. In practice, they may correspond to doctors and hospitals, students and colleges, teachers and schools, men and women, actual firms and workers for labor markets with fixed wages (e.g., government and union jobs, see Hall and Krueger, 2012), or any other matching setting with limited or no transfers at all.

[17] Our main results do not rely on all agents being acceptable and can be extended.

[18] Experimental studies suggest that subjects have a limited degree of sophistication even in relatively small decentralized markets. Additionally, subjects predominantly form successive blocking pairs (Echenique, Robinson-Cortés, and Yariv, 2023; Pais, Pintér, and Veszteg, 2020).

Formally, let matching λ be unstable. For any blocking pair (f_i, w_j) of λ, a matching μ is *obtained from* λ *by satisfying* (f_i, w_j) if (i) (f_i, w_j) are matched to each other in μ; (ii) their partners at λ, if any, are unmatched at μ; (iii) and all other agents are matched identically under both μ and λ. In other words, we match agents inside the blocking pair, divorce their original partners, and maintain matches of all other agents. We say that there is a *path* (of *length* $k-1$) from matching λ to matching μ if there exists a sequence of matchings $\lambda_1, \lambda_2, \ldots, \lambda_k$, such that $\lambda = \lambda_1$, $\mu = \lambda_k$, and for each $i < k$, λ_{i+1} is obtained from λ_i by satisfying a blocking pair. In that case, we say matching λ can *reach* or *attain* matching μ.[19]

Theorem (Roth and Vande Vate, 1990)**.** *For any unstable matching, there is a finite sequence of blocking pairs that leads to a stable matching.*

Consider a random decentralized process that starts at matching $\lambda = \lambda_1$ and successively produces matchings λ_i in such a way that for any $i > 1$, matching λ_{i+1} is obtained from λ_i by satisfying a single blocking pair, chosen randomly from all possible blocking pairs for λ_i. Additionally, suppose that at each step, for any two blocking pairs, one blocking pair is at most $\kappa \geq 1$ times as likely to be satisfied than the other.[20] We refer to such a random process as *κ-random dynamics* for brevity. Then,

Corollary (Roth and Vande Vate, 1990)**.** *For any unstable matching, the sequence of blocking pairs, produced by κ-random dynamics, converges to a stable matching with certainty.*

The example below illustrates the dynamics of blocking-pair formation in a simple market.

Example 1. Consider a market with two firms and two workers. The following payoff matrix describes the agents' ordinal preferences in a convenient way:

$$\begin{array}{c} \\ f_1 \\ f_2 \end{array} \begin{array}{c} \begin{array}{cc} w_1 & w_2 \end{array} \\ \begin{pmatrix} 2,1 & 1,2 \\ 1,2 & 2,1 \end{pmatrix} \end{array}.$$

In this matrix notation, (i) each row i corresponds to firm f_i and each column j corresponds to worker w_j; (ii) every (i,j)-entry specifies payoffs for f_i and w_j, respectively, if they are

[19] This process originates from Knuth (1976). Knuth assumes an equal number of firms and workers and requires divorced agents to match when satisfying a blocking pair. He showed that such a process may cycle and raised a question if we can always find a path to stability. More than a decade later, Roth and Vande Vate (1990) provided a positive answer for the setting used in the current paper. In Knuth's setting, though, sometimes it is impossible to attain stability (Tamura, 1993; Tan and Su, 1995).

[20] Equivalently, for any unstable matching, whenever it arises, each of its blocking pairs is satisfied with a probability bounded away from zero as i goes to infinity; see footnote 8 in Roth and Vande Vate (1990).

matched with each other; (iii) larger payoffs correspond to more preferred partners; (iv) and all payoffs from being unmatched are normalized to zero. For instance, f_2 finds all workers acceptable and prefers w_2 to w_1.

There are two stable matchings. Matching μ_F in which agents with identical indices are matched, $\mu_F(f_i) = w_i$, is the firm-optimal stable matching: each firm weakly prefers her assigned partner to the partner she would get in any other stable matching. The worker-optimal stable matching μ_W pairs agents with different indices, $\mu_W(f_i) = w_j$ for $i = 3 - j$.[21]

It is easy to verify that starting from any unstable matching, the decentralized process can attain any stable matching. For instance, consider unstable matching λ obtained from the firm-optimal stable matching μ_F by unmatching firm f_2 with her stable partner, worker w_2; that is, $\lambda(f_1) = w_1$ and $\lambda(f_2) = f_2$. Then, there exists a path, of length two, from this "almost firm-optimal stable" matching λ to the worker-optimal stable matching μ_W:

$$\lambda = \lambda_1 \xrightarrow[(f_2,w_1)]{} \lambda_2 = \begin{matrix} & w_1 & w_2 & \\ & | & | & \\ (& f_2, & w_2 &) \end{matrix} \xrightarrow[(f_1,w_2)]{} \lambda_3 = \mu_W,$$

where under each transition arrow, we specify the corresponding blocking pair to be satisfied.

Suppose that at each step, a blocking pair is selected uniformly at random; this corresponds to 1-random dynamics. Then, λ converges to μ_W with probability

$$p = \underbrace{1/2}_{(f_2,w_2)} \times 0 + \underbrace{1/2}_{(f_2,w_1)} \times (1-p) \quad \Longrightarrow \quad p = 1/3.$$

Indeed, there are two equally-likely blocking pairs for matching λ, (f_2, w_2) and (f_2, w_1). By satisfying pair (f_2, w_2), we end up in the firm-optimal stable matching μ_F. The process terminates, with zero probability of convergence to μ_W; see the first term. Otherwise, if we satisfy blocking pair (f_2, w_1), we obtain "almost worker-optimal stable" matching λ_2. This matching converges to μ_F with probability p due to symmetry, leaving the remaining probability of $1 - p$ for convergence to μ_W; see the second term. $\triangle$

In our paper, we also explore a variation of the Roth and Vande Vate (1990) dynamics, wherein agents successively select their *best* blocking partners. Specifically, a blocking pair is the *best blocking pair* for an agent if she prefers her partner in this pair to any other partner with whom she could form a blocking pair. Analogously to κ-random dynamics, we define *κ-random best dynamics* by substituting "blocking pairs" with "*best* blocking pairs" in the definition of κ-random dynamics. In fact, due to Ackermann et al. (2011), for any

[21] See Appendix A for further details on the structure of stable matchings.

unstable matching, there exists a finite sequence of best blocking pairs that leads to a stable matching. Consequently, results similar to the above-mentioned theorem and corollary hold when we restrict attention to best blocking pairs.

Both versions of random dynamics particularly allow for non-uniform and time-dependent formation of (best) blocking pairs. The probability that a specific blocking pair of agents matches at a given time can be quite general, reflecting factors such as the likelihood that these agents would encounter each other (Roth and Vande Vate, 1990). Alternatively, the probability may depend on the agents' incentives to form a match. One may assume, for instance, that agents have cardinal payoffs, and a blocking pair generating a higher total surplus—or a higher total surplus gain compared to the former match—is more likely to be formed. We present our main results under these broad classes of decentralized dynamics.

These random dynamics capture the inherently stochastic and sequential nature of decentralized interactions; this by no means implies that agents act without intention or reason. The randomness reflects the complexity and unpredictability of the environment, where agents, driven by a desire for improvement, continuously explore and adapt to evolving opportunities amidst uncertainty about future interactions. It implicitly captures the complex factors and considerations guiding agents, leading to uncertainty about which pair of agents will match next. As each match forms, it closes some opportunities while potentially creating new ones, guiding the market's evolution.

3. Anything Goes

In this section, we prove that, under mild conditions, starting from *any* unstable matching, decentralized interactions can attain *any* stable matching. For simplicity, we consider balanced markets, with an equal number n of agents on both sides, for the rest of the paper.[22]

Formally, consider a subset of firms and a subset of workers of equal size $k < n$. They form a submarket with preferences inherited from the original market. Suppose that this submarket has a stable matching such that any agent—firm or worker—inside the submarket prefers her stable partner under that matching to every agent outside the submarket. In this case, we say that the above two subsets constitute a *fragment*. A fragment is *trivial* if all stable matchings in the original market agree on how they match agents inside the fragment.

In addition, we say that a matching is *almost stable* if it has a path of length one to a stable matching; that is, it is just one blocking pair away from stability.[23] Notably, the

[22] All definitions and results can be generalized in a straightforward manner to imbalanced markets.

[23] This concept is partially related to *almost stable matchings* of Abraham, Biró, and Manlove (2005) and

theorem below holds even when we restrict attention to best blocking pairs.

Theorem 1. *The following three statements are equivalent:*

(i) for any unstable matching λ and any stable matching μ, there exists a finite sequence of (best) blocking pairs that leads from λ to μ;

(ii) for any almost stable matching λ and any stable matching μ, there exists a finite sequence of (best) blocking pairs that leads from λ to μ;

(iii) there are no non-trivial fragments.

This result shows one type of fragility of stable matchings. Small perturbations of a stable matching may yield *any* stable matching, close to or distant from the perturbed matching.[24]

Corollary 1. *Suppose that there are no non-trivial fragments. For any unstable matching, the sequence of (best) blocking pairs, produced by κ-random (best) dynamics, converges to any stable matching with positive probability.*

This corollary is further refined in Sections 4 and 5, illustrating that decentralized interactions often lead the market away from a slightly perturbed stable matching, with no guarantee of returning or even converging to a close, similar stable matching. Such "local instability" of stable matchings suggests that decentralized interactions can significantly reshape or reverse market outcomes, even after minor deviations from the targeted stable outcome in a centralized market.

In what follows, we discuss fragments in more detail and elucidate their role in constraining decentralized interactions. We argue that the absence of non-trivial fragments is a mild condition. We then present a sketch of the proof of the theorem.

---

minimally unstable matchings of Doğan and Ehlers (2021). These papers consider settings in which a stable matching may not exist and focus on matchings that are "as stable as possible."

[24]This result holds for an *arbitrary* starting matching. If the starting matching is empty with all agents single, decentralized interactions are unrestrained, and any stable matching can be attained by sequentially pairing stable pairs constituting that matching. Relatedly, a two-sided version of the deferred acceptance algorithm (Dworczak, 2021), which starts from the empty matching and allows both firms and workers to propose, can reach any stable matching; this result and its proof, like the previous observation, crucially rely on the initial matching being *empty*. In contrast, we consider *arbitrary* deviations from stability, as our conceptual focus is on the robustness of stable matchings rather than on algorithms to find them. In doing so, we introduce the concept of fragments and uncover their unique role in restricting interactions.

3.1. Fragments. Consider subsets $\bar{F} \subsetneq F$ and $\bar{W} \subsetneq W$ of equal size $k < n$. Let $\bar{\mathcal{M}}$ be the market induced by the original market $\mathcal{M}$ when restricted to $\bar{F} \times \bar{W}$. Firms $\bar{F}$ and workers $\bar{W}$ form a *fragment*, of *size* k, in $\mathcal{M}$ if there exists a stable matching $\bar{\mu} : \bar{F} \cup \bar{W} \to \bar{F} \cup \bar{W}$ for $\bar{\mathcal{M}}$ such that (i) for any $\bar{f} \in \bar{F}$ and $w \notin \bar{W}$, $\bar{\mu}(\bar{f}) \succ_{\bar{f}} w$, (ii) and for any $\bar{w} \in \bar{W}$ and $f \notin \bar{F}$, $\bar{\mu}(\bar{w}) \succ_{\bar{w}} f$. In this case, we say that matching $\bar{\mu}$ *induces* fragment $(\bar{F}, \bar{W})$.[25]

Fragments have the following defining property: any stable matching in the original market must match agents inside a fragment within the fragment. That is why we call them "fragments." Nonetheless, stable matchings may still disagree with each other, and with all inducing matchings, on how they match agents inside the fragment.

Lemma 1. *Suppose that firms $\bar{F}$ and workers $\bar{W}$ form a fragment. Then, $\mu(\bar{F}) = \bar{W}$ for for any stable matching μ in the original market.*

Intuitively, consider the fragment induced by matching $\bar{\mu}(f_i) = w_i$, $i \leq k < n$. Assume by contradiction that, say, worker w_1 is not matched within the fragment for some stable matching μ. Then, he prefers his partner under $\bar{\mu}$, firm f_1. By the stability of μ, firm f_1 prefers her partner under μ to worker w_1. Therefore, by the definition of a fragment, f_1 must be matched within the fragment under μ, say, with worker w_2. By the stability of $\bar{\mu}$ for the fragment, w_2 prefers his stable partner under $\bar{\mu}$, firm f_2, to firm f_1. This implies that f_2 prefers her partner under μ to worker w_2, again by the stability of μ. Hence, f_2 also must be matched within the fragment under μ, say, with worker w_3. We can proceed iteratively until we reach the last firm, firm f_k, that cannot be matched within the fragment since all workers inside the fragment are already matched with other firms. This yields a contradiction.

A fragment is called *trivial* if all stable matchings in the original market agree on how they match agents within the fragment; this is more restrictive than the conclusion of Lemma 1. If a fragment is trivial, it has the unique inducing matching, and all stable matchings in the original market coincide with that inducing matching on the fragment. For instance, a firm-worker pair that are each other's favorite partner constitutes a trivial fragment of size one. This pair is called a *top-top match*. Furthermore, a sequence of $k < n$ such pairs—where each new top-top pair in the sequence is obtained from the submarket after removing all previous top-top pairs from the original market—forms a trivial fragment of size k. Nonetheless,

[25] A market may have multiple fragments: nested, overlapping, or disjoint. In addition, a fragment might be induced by multiple matchings; see Example 3 in the Online Appendix. For any inducing matching, we can find a stable matching in the original market that agrees with that inducing matching over its corresponding fragment, as shown by Lemma 4 in the Online Appendix. The converse is not true. That is, a stable matching in the original market might disagree with all matchings inducing a fragment—in Example 2 below, the worker-optimal stable matching μ_W disagrees with the unique inducing matching $\bar{\mu}$.

trivial fragments are not limited to such sequences, see Example 4 in the Online Appendix. The following example shows how a *non-trivial* fragment might restrain the dynamics.

Example 2. Consider a market with three firms and three workers

$$\begin{array}{c} \\ f_1 \\ f_2 \\ f_3 \end{array} \begin{array}{c} \begin{array}{ccc} w_1 & w_2 & w_3 \end{array} \\ \begin{pmatrix} \mathbf{3,2} & 1,3 & 2,1 \\ 1,3 & \mathbf{3,2} & 2,2 \\ 3,1 & 2,1 & 1,3 \end{pmatrix} \end{array}$$

that has two stable matchings, $\mu_F = (f_1, f_2, f_3)$ and $\mu_W = (f_2, f_1, f_3)$. As a side note, for this market and other markets with more than two workers, we use a simplified matching notation that specifies partners of workers w_1, w_2, and so forth. For instance, $\mu_W = (f_2, f_1, f_3)$ means that w_1 is matched with f_2, w_2 with f_1, and w_3 with f_3.

Firms $\bar{F} = \{f_1, f_2\}$ and workers $\bar{W} = \{w_1, w_2\}$ form a fragment induced by matching $\bar{\mu}$ that couples firms f_1 and f_2 with workers w_1 and w_2, respectively. It is non-trivial since the worker-optimal stable matching μ_W disagrees with the inducing matching $\bar{\mu}$ on the fragment.

In contrast to Example 1, some unstable matchings cannot attain all stable ones. Consider almost stable matching $\lambda = (f_1, f_2, w_3)$ that agrees with the inducing matching $\bar{\mu}$ on the fragment. Matching $\bar{\mu}$ traps the decentralized process. Indeed, since $\bar{\mu}$ is stable for the fragment, no agent inside it can form a blocking pair with anyone that is also inside. Also, by definition, every agent inside the fragment prefers her partner under $\bar{\mu}$ to every agent outside it. Altogether, no agent inside the fragment can form a blocking pair with anyone, neither inside nor outside it. Thus, the dynamics cannot reach any stable matching that disagrees with $\bar{\mu}$. Non-triviality ensures that such a stable matching—here, μ_W—exists. $\triangle$

We argue that a non-trivial fragment is demanding and most markets have no such fragments. A non-trivial fragment requires that (i) agents inside the fragment are matched in a *stable* way; (ii) they prefer their stable partners to *everyone* outside; (iii) and there is *another stable* way to match these agents in the original market. As an illustration, consider markets with uniformly random preferences. Numerical simulations show that, with high probability, there are no non-trivial fragments at all; we provide theoretical results regarding probabilistic aspects of fragments in the Online Appendix.[26] Figure 1 illustrates

[26]Fragments are a novel concept, and their analysis appears rather delicate. The Online Appendix presents results on the likelihood of fragments and discusses associated challenges. Specifically, we demonstrate that, with high probability, there are no fragments at all, even trivial ones, that involve a large fraction of agents. We also introduce techniques, in the spirit of Pittel (1989), that may be potentially useful for future research.

that, although the frequency of markets with non-trivial fragments is increasing for market sizes of up to $n = 7$, this frequency never exceeds 20% and eventually decreases with the market size.[27] Thus, for such markets, Theorem 1 holds almost generically.[28]

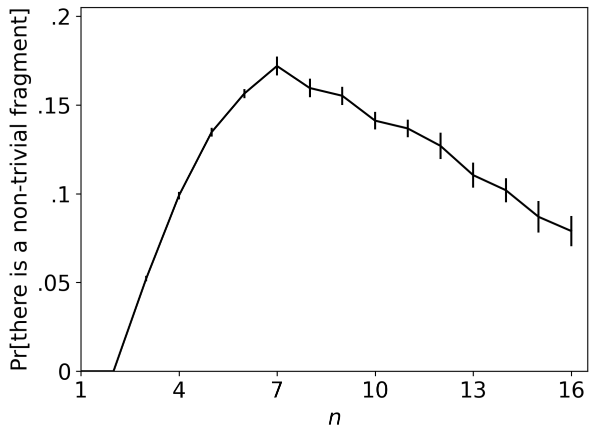


Figure 1: Frequency of random markets with non-trivial fragments

Furthermore, even when non-trivial fragments are present, they typically restrain the dynamics only for a small fraction of specific unstable matchings. The great majority of unstable matchings can still attain many, if not all, stable matchings.[29]

3.2. Sketch of the Proof. In the proof of the theorem, we use one more piece of notation. Let μ_{-f} denote the almost stable matching obtained from stable matching μ by unmatching firm f with her stable partner, worker $\mu(f)$.

We prove the theorem by establishing the following cycle of implications: $(ii) \Rightarrow (i) \Rightarrow (iii) \Rightarrow (ii)$. To avoid technicalities, we provide a sketch of the proof for the case when no restriction is placed on blocking pairs; the full proof is relegated to Appendix B. The first implication $(ii) \Rightarrow (i)$ follows from the stabilization theorem of Roth and Vande Vate (1990). Specifically, consider any unstable matching λ. By the stabilization result, there is a finite

[27] We simulate $n \times n$ markets, the number of which varies from 1,000 for larger n to 20,000 for smaller n.

[28] At the same time, balanced markets with uniformly random preferences have many stable matchings—all of which are anticipated to be fragile—with significantly different welfare properties. Formally, Pittel (1989) shows that, asymptotically, the expected number of stable matchings is $e^{-1} n \ln n$. Moreover, the firms' average ranks of their assigned workers under the firm-optimal ($\ln n$) and worker-optimal ($n / \ln n$) stable matchings are substantially different; analogous results hold for workers. The above exercise becomes less relevant for imbalanced markets with uniformly random preferences (Ashlagi, Kanoria, and Leshno, 2017).

[29] Additionally, if a non-trivial fragment is considered as a separate market and itself is free of other non-trivial fragments, our insights still apply to this market as well.

sequence of blocking pairs that leads to some stable matching μ:

$$\lambda = \lambda_1 \longrightarrow \lambda_2 \longrightarrow \ldots \longrightarrow \lambda_{k-1} \longrightarrow \mu = \lambda_k.$$

Therefore, λ can reach almost stable matching λ_{k-1}, from which in turn it can attain any stable matching by the premise (ii).

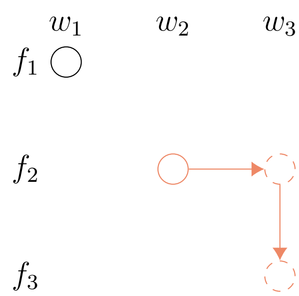
The second implication $(i) \Rightarrow (iii)$ holds by the definition of a non-trivial fragment. Suppose by contradiction that there is a non-trivial fragment. Consider an unstable matching that, when restricted to the fragment, agrees with its inducing matching. Then, this unstable matching cannot yield any stable matching that instead disagrees with the inducing matching over the fragment. Since the fragment is non-trivial, such a stable matching exists, thus contradicting the premise (i); this argument is similar to Example 2.

Finally, and most importantly, why can decentralized interactions lead to any stable matching if there are no non-trivial fragments? The third implication $(iii) \Rightarrow (ii)$ is the key part of the proof. We prove this implication by induction on the size n of matching markets with no non-trivial fragments. The result is trivial for $n = 1$, when there is only one agent on each side of the market. It also holds for such markets of size $n = 2$. This is because it is valid for the market from Example 1 that is the only 2×2 market with multiple stable matchings, up to permutations of firms' and workers' indices.

To glean intuition behind the induction step, assume that the result holds for all markets with no non-trivial fragments of size up to $n-1$ and focus on any such market of size n. Consider any almost stable matching λ which corresponds to some stable matching μ. The proof consists of two main observations. The first observation is at the core of the proof and crucially relies on the properties of fragments as well as the induction hypothesis. We show that λ can reach either any stable matching—as desired—or any other almost stable matching for stable matching μ. Hence, suppose that any almost stable matching for μ can be attained. The second observation, which we also describe in more detail later in this section, is more technical. This observation connects our procedure to the deferred acceptance algorithm in order to construct a path from one of those almost stable matchings to a new almost stable matching corresponding to some new stable matching. We employ these two observations repeatedly until we find paths to all stable matchings.

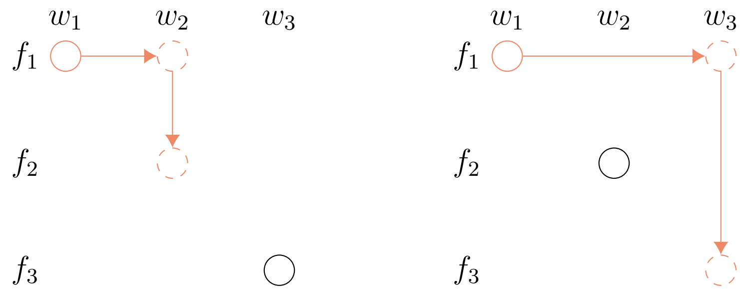
To illustrate the first observation, consider any market with no non-trivial fragments of size $n = 3$ and an arbitrary almost stable matching λ. Up to relabeling, almost stable $\lambda = \mu_{-f_3}$ is obtained from stable matching $\mu = (f_1, f_2, f_3)$ by unmatching firm f_3 with her stable partner, worker w_3. We show that μ_{-f_3} can attain either any stable matching or any other almost stable matching μ_{-f_i}, $i < 3$.

Focus on the subsets $\bar{F} = \{f_1, f_2\}$ and $\bar{W} = \{w_1, w_2\}$ of firms and workers, respectively. Suppose first that these subsets form a fragment. This fragment must be trivial by the premise. Therefore, all stable matchings match f_1 with w_1 and f_2 with w_2. This implies that matching μ is the unique stable matching, so the result follows immediately, say, by the stabilization theorem. Henceforth, suppose that the subsets $\bar{F}$ and $\bar{W}$ do not constitute a fragment. Since stable matching μ remains stable when restricted to the submarket induced by these subsets, there is an agent inside these subsets that prefers someone outside to her stable partner under μ. By symmetry, let firm f_2 prefer worker w_3 to her stable partner, worker w_2. By the stability of matching μ in the original market, w_3 must prefer his stable partner, firm f_3, to firm f_2. But then almost stable matching μ_{-f_2} can be attained by successively satisfying blocking pairs (f_2, w_3) and (f_3, w_3). This can be described by the following diagram in which solid circles represent initial matches:



Thus, μ_{-f_3} can attain the new almost stable matching μ_{-f_2}.

We next consider smaller subsets of firms $\bar{F} = \{f_1\}$ and workers $\bar{W} = \{w_1\}$. Suppose first that they form a fragment. It must be trivial by the premise; in fact, top-top match (f_1, w_1) is a trivial fragment by definition. Thus, all stable matchings match f_1 with w_1. As it turns out, after removing agents forming a trivial fragment from the original market, the remaining market—that consists of firms $F \setminus \bar{F}$ and workers $W \setminus \bar{W}$—continues to have no non-trivial fragments (Lemma 3 in Appendix B). When restricted to the remaining market, μ_{-f_3} is unstable and, by the induction hypothesis, can reach any stable matching in this smaller market. Since all stable matchings in the original market agree on the removed trivial fragment and persist to be stable in the remaining market, matching μ_{-f_3} can reach all stable matchings in the original market, as desired. Henceforth, assume that $\bar{F}$ and $\bar{W}$ do not form a fragment. As before, by symmetry, let firm f_1 prefer some worker outside, w_2 or w_3, to her stable partner, worker w_1. In either of these two cases, the stability of matching μ ensures that matching μ_{-f_1} can be attained in two steps from either μ_{-f_2} or μ_{-f_3}:



Consequently, μ_{-f_3} can reach any other almost stable matching μ_{-f_i}, $i < 3$, corresponding to μ. This concludes the first observation.

In the second observation, we connect the decentralized process of our interest to the deferred acceptance algorithm. McVitie and Wilson (1971) adapt this algorithm by introducing a new operation to determine the set of all stable matchings. Their adapted version can be roughly described as follows; see the full details and relevant results in Appendix A. They first run the standard firm-proposing version to compute the firm-optimal stable matching μ_F, the best stable matching for firms. Next, they seek to obtain any other stable matching. Particularly, fix an arbitrary stable matching $\mu \neq \mu_F$. Then, some firm f prefers worker $w = \mu_F(f)$ to worker $\mu(f)$. McVitie and Wilson (1971)'s idea is to break the (f, w)-partnership and restart the previously terminated deferred acceptance algorithm by forcing firm f to propose to the worker following w in her list. As it turns out, this operation called *breakmarriage* generates a new stable matching μ' that is at least as good as matching μ for all firms. If μ and μ' are different, some firm f' prefers worker $w' = \mu'(f')$ to worker $\mu(f')$. McVitie and Wilson (1971) then break the (f', w')-partnership, restart the deferred acceptance algorithm one more time, and proceed iteratively. They show that any stable matching can be obtained by successive applications of breakmarriage operations. In Lemma 2 in Appendix A, we prove that our dynamics can replicate the restarted deferred acceptance algorithm following breakups for all relevant breakmarriage operations.

We conclude the proof by combining the two observations. The first observation suggests that the absence of non-trivial fragments effectively allows to attain any almost stable matching for a particular stable matching. Equivalently, we can break any stable partnership in that stable matching. This, together with the second observation, implies that all relevant breakmarriage operations in McVitie and Wilson (1971)'s algorithm can be emulated by the dynamics. We use this connection to establish the result.

4. Exponential Stabilization

We now turn to another form of fragility. Even though convergence to stability is always guaranteed, the market may be far from stable for long periods of time. To isolate this form of fragility, this section examines markets with a unique stable matching. In such markets, the dynamics necessarily converge to one matching, the unique stable one. As mentioned earlier, our focus on markets with a unique stable matching is also in line with the literature suggesting that large markets entail small cores.[30] Henceforth, we analyze the random dynamics introduced in Section 2.2, which allow for non-uniform and time-dependent sequential formation of (best) blocking pairs.

Our second theorem shows that for a large class of markets with a unique stable matching, even a small deviation from stability leads to an exponentially long stabilization process. More precisely, we consider *any* market with a unique stable matching. We prove that the market can be augmented by adding a small fraction of new agents so that the resulting market still has a unique stable matching, but where even for small perturbations of the stable matching, stabilization takes an exponentially long time.[31]

Formally, a δ-*augmented market* of the original market of size n is a new market of size at most $(1+\delta)n$ obtained from the original market by adding at most a δ-proportion of new agents on both sides, so that (i) both the original and the augmented markets coincide when restricted to the original agents; (ii) the augmented market has a unique stable matching; (iii) and all original agents have identical stable partners in both markets. Importantly, an augmentation does not alter the original market and its structure of stable matchings.

We are interested in small deviations from stability. A matching is ϵ-*unstable* if at least an ϵ-fraction of firms, and thus workers, are not matched with their unique stable partners.

Theorem 2. *Focus on κ-random (best) dynamics. Consider any sequence of markets of size $n \in \mathbb{N}$ with a unique stable matching and any $\delta > 0$, $\epsilon \in (0,1]$, and $\kappa \geq 1$. Then, there exists a corresponding sequence of δ-augmented markets for which any sequence of ϵ-unstable matchings with probability $1 - 2^{-\Omega(n)}$ needs $2^{\Omega(n)}$ steps to regain stability.*[32]

[30] Specifically, theoretical work has identified various conditions under which in large markets, all agents, except for a vanishing proportion, are matched with the same partners in all stable matchings; see Immorlica and Mahdian (2005), Kojima and Pathak (2009), and Ashlagi, Kanoria, and Leshno (2017), as well as our discussion in footnote 3. Although our first theorem applies to such markets and suggests re-equilibration dynamics, it becomes less relevant since all stable matchings are essentially identical.

[31] Our approach is reminiscent on an abstract level of the "maximal domain" exercises common in the matching literature (e.g., see Gul and Stacchetti, 1999; Hatfield and Kominers, 2012; Kamada and Kojima, 2023) and similar in spirit to the augmentation exercises in Fernandez, Rudov, and Yariv (2022).

[32] We write $f(n) = \Omega(g(n))$ if $g(n) = O(f(n))$.

This theorem emphasizes an additional fragility aspect of stable matchings; even for slight deviations from stability, the stabilization process takes an extremely long time. A sketch of the proof of the theorem is provided in Section 4.1. At a very high level, we construct augmented markets in which, for many matchings along any path to stability, there are substantially more "destabilizing" (best) blocking pairs that move the market away from stability than "stabilizing" (best) blocking pairs—corresponding to stable pairs of agents—that instead facilitate the convergence. This significantly slows down the convergence and causes exponentially long stabilization paths.[33,34]

Theorem 2, along with its underlying intuition, naturally complements Theorem 1. In fact, the simulations in Section 5 demonstrate that for random markets, irrespective of whether stable matchings are unique, and even when initializing a market at almost stable matchings, the stabilization process typically leads the market away from stability and takes a very long time to regain it.[35] Due to the prevalence of "destabilizing" rematching opportunities, many market participants are mismatched for extended periods. Furthermore, for markets with multiple stable matchings, even almost stable matchings converge with substantial probability to a stable matching distant from the minimally perturbed matching; this further refines our findings in Section 3, particularly Corollary 1.

In the rest of this section, we first outline the proof of the second theorem. We then discuss one more way in which fragments can restrain decentralized interactions.

4.1. Sketch of the Proof. The proof proceeds in two steps. First, we establish an analogous result for a class of non-augmented markets having a unique stable matching. Then, we

[33] In our analysis of Theorem 2, we focus on random dynamics that guarantee convergence to stability, consistent with our discussion in Section 2.2. To produce long paths—though not necessarily leading to stability—an even broader class of dynamics suffices, wherein "destabilizing" pairs are not vastly less likely to be satisfied compared to other pairs. Moreover, it seems feasible to allow $\kappa = \kappa(n)$ to increase with n, albeit at the expense of introducing unnecessary complexity into the analysis.

[34] While specific paths to stability may be polynomially long, such paths are vanishingly rare. This applies to paths produced by a two-sided version of the deferred acceptance algorithm with compensation chains (Dworczak, 2021), where agents dropped by partners who had made them offers initially get priority in proposing themselves. As Dworczak (2021) notes, the inclusion of compensation chains in the already structured deferred acceptance algorithm serves the sole theoretical purpose of guaranteeing that the algorithm terminates at a stable matching. These algorithm-driven restrictions not only lead to unlikely paths but also render the dynamics underlying this algorithm not representative of decentralized settings, where interactions are inherently less structured and predictable (Echenique et al., 2023).

[35] Obtaining this result theoretically is challenging due to two layers of randomness inherent in both random markets and random decentralized dynamics. Addressing this challenge would require not only alternative methods but also specific assumptions about randomness. Instead, similar to our first theorem, Theorem 2 applies to non-random markets, specifically to all markets having a unique stable matching. Moreover, our result holds for a broad range of (potentially non-uniform and time-dependent) random dynamics.

utilize markets from this class to augment arbitrary markets and demonstrate the theorem.

For the first step, consider a market of size $n \in \mathbb{N}$ with a unique stable matching μ, where for some $\eta \in (0,1)$, the following two conditions hold: (i) any firm *except* firm $\bar{f}$ is preferred by at least an η-proportion of workers to their stable partners; (ii) and similarly, any worker *except* worker $\bar{w}$ is preferred by at least an η-proportion of firms to their stable partners. These exceptions arise because in a market with a unique stable matching, it is impossible for *every* agent on one side to be preferred by some agents from the other side to their stable partners.[36] We prove that any sequence of such markets (indexed by sizes $n \in \mathbb{N}$), where η is fixed along the sequence, achieves a conclusion parallel to that of the theorem; see Proposition 1 in Appendix C. The identified class of markets, which is non-empty and quite broad, particularly includes certain "almost" fully assortative markets, wherein all agents on each side agree on how they rank the vast majority of all agents from the other side.

To illustrate the idea behind Proposition 1, consider the market of size n described earlier. For any matching λ, let $\mathcal{S}(\lambda)$ denote the number of firms, and hence workers, matched with their stable partners. In order to restore stability, $\mathcal{S}(\mu) = n$, at some point of the process, the market must enter the region $\mathcal{S}(\lambda) \geq (1-\zeta)n$, where $\zeta \leq \epsilon$, and remain in this region until convergence; in fact, $\zeta > 0$ can be arbitrarily small. Therefore, consider any unstable matching λ with $\mathcal{S}(\lambda) \geq (1-\zeta)n$. To avoid technical difficulties, suppose λ satisfies the following condition ($\star$): some agent $a \notin \{\bar{f}, \bar{w}\}$ is unmatched. In that case, this agent is preferred by at least an η-proportion of agents from the other side to their stable partners. Since at least a $(1-\zeta)$-fraction of agents are matched with their stable partners in matching λ, there are at least $(\eta - \zeta)n$ destabilizing (best) blocking pairs that generate matchings λ' with $\mathcal{S}(\lambda') = \mathcal{S}(\lambda) - 1$; note, however, that new matchings λ' might fail to satisfy ($\star$). At the same time, there are at most ζn stabilizing (best) blocking pairs that lead to matchings λ' with $\mathcal{S}(\lambda') = \mathcal{S}(\lambda) + 1$; they may not satisfy ($\star$) as well. For small enough $\zeta > 0$, there are many more destabilizing pairs than stabilizing ones. If all matchings along the process satisfied ($\star$), we could associate the dynamics with a random walk that is heavily biased in the "destabilizing" direction. Such random walks are known to take exponentially many steps, with high probability, to reach $\mathcal{S}(\mu) = n$. In our proof, since some matchings resulting from the dynamics violate ($\star$), we use more sophisticated arguments to establish the result.

In the second step of the proof, we can use any sequence of constructed markets from the identified class, for which Proposition 1 holds, to δ-augment the original markets and establish the theorem by analogous techniques. For each original market, of size n, add

[36] See Lemma 7 in Balinski and Ratier (1997).

new agents with preferences over each other that mirror the constructed market of size $\lfloor \delta n \rfloor$. Additionally, let any agent, original or new, prefer original agents to new ones. The resulting market, of size $\lfloor \delta n \rfloor$, is indeed δ-augmented and satisfies conditions akin to those described in Proposition 1. Despite the sizes of the δ-augmented markets being $n + \lfloor \delta n \rfloor$ rather than n, they still increase linearly with $n \in \mathbb{N}$. The desired exponential convergence result then follows by similar arguments as before, completing the proof of the theorem.

4.2. Fragments Revisited. Section 3, and notably Theorem 1, sheds light on the special role of non-trivial fragments in constraining the stabilization dynamics. Below, we illustrate another way in which fragments, particularly trivial ones, restrain decentralized interactions.

Specifically, consider a market in which firms and workers can be relabeled so that subsets $\bar{F}_1 = \{f_1\}$ and $\bar{W}_1 = \{w_1\}$ form a fragment of size one, $\bar{F}_2 = \bar{F}_1 \cup \{f_2\}$ and $\bar{W}_2 = \bar{W}_1 \cup \{w_2\}$ constitute a fragment of size two, and so forth. In particular, firms $\bar{F}_{n-1} = \bar{F}_{n-2} \cup \{f_{n-1}\} = \{f_1, f_2, \dots, f_{n-1}\}$ and workers $\bar{W}_{n-1} = \bar{W}_{n-2} \cup \{w_{n-1}\} = \{w_1, w_2, \dots, w_{n-1}\}$ generate a fragment of size $n-1$. In other words, a market has a *nested structure of fragments*; all these fragments are trivial by definition. This market can also be equivalently described as a sequence of top-top match pairs. Firm-worker pair (f_1, w_1) is a top-top match: they are each other's favorite partner. Furthermore, for any $i \geq 2$, pair (f_i, w_i) is a top-top match in the submarket formed by firms $F \setminus \bar{F}_{i-1}$ and workers $W \setminus \bar{W}_{i-1}$. This alternative formulation coincides with the *sequential preference condition*, first introduced by Eeckhout (2000).

Any such market has a unique stable matching that can be derived by successively partnering top-top match pairs. Many matching markets studied in the literature have a nested structure of fragments.[37] This class includes assortative markets, in which agents on one or both sides share the same ranking of agents from the other side.

A nested structure of fragments makes the stabilization dynamics polynomial. Consider an arbitrary unstable matching and, as an example, focus on 1-random dynamics: at each step, a blocking pair is selected uniformly at random. If firm f_1 and worker w_1 from top-top match pair (f_1, w_1) are not yet matched with each other, this pair remains a blocking pair until they are matched. Since there are at most n^2 blocking pairs at each step of the dynamics, they are expected to match in $O(n^2)$ time. Once matched, f_1 and w_1 continue to be partners until convergence, which means the decentralized process gets stuck in fragment $(\bar{F}_1, \bar{W}_1)$. Next, if firm f_2 and worker w_2 from top-top match pair (f_2, w_2) are not yet

[37]For instance, the α-reducibility property (Clark, 2006), the aligned preferences condition (Ferdowsian, Niederle, and Yariv, 2023), and oriented preferences (Reny, 2021) satisfy Eeckhout (2000)'s condition and thus generate a nested structure of fragments.

matched, this pair remains a blocking pair until it is resolved. These agents are expected to match with each other in $O(n^2)$ time as well. The dynamics then get stuck in fragment $(\bar{F}_2, \bar{W}_2)$. By iterative arguments, the stabilization takes $O(n^3)$ expected time.[38]

Nonetheless, even markets in which the convergence is polynomial are potentially fragile, as demonstrated by Theorem 2. A slight proportion of agents can be added to make the stabilization exponential. In fact, the construction in the theorem's proof can be straightforwardly modified to ensure that augmented markets have no fragments at all.

5. Computational Experiments

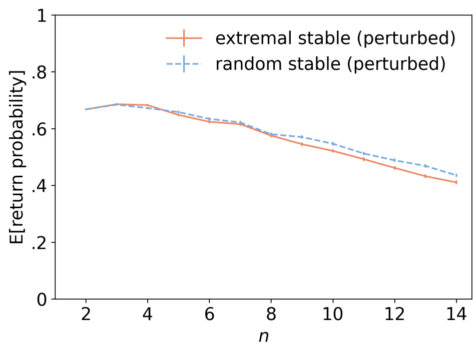
This section presents simulation results for random markets that complement and sharpen our theoretical findings. In markets with multiple stable matchings, even a minimal perturbation of a stable matching often converges to a stable matching distant from the perturbed matching. Furthermore, regardless of whether stable matchings are unique, the stabilization dynamics typically stray from stability, take a long time to reach stability, and involve a sizable proportion of mismatched agents for long durations. For simplicity, suppose that at each step of the dynamics, a blocking pair is chosen uniformly at random.[39]

First, we consider markets with multiple stable matchings. For each market size n, we simulate $n \times n$ markets with uniformly random preferences and select 1,000 markets with multiple stable matchings. In each of these markets, we examine two stable matchings. One is an extremal stable matching, which is often targeted by centralized clearinghouses and is thus of special interest; by symmetry, the analysis is identical for firm- and worker-optimal stable matchings. The other is chosen uniformly at random from the set of all stable matchings, including extremal stable matchings. To provide a lower bound on the fragility of stable matchings, we focus on minimal deviations from a stable matching. Specifically, for each of the two stable matchings, we select one almost stable matching uniformly at random. Next, we calculate various statistics, such as the probability of returning to the minimally perturbed stable matching and the expected time to regain stability, by simulating 300 paths to stability. Then, for each market size and each perturbed stable matching, we compute

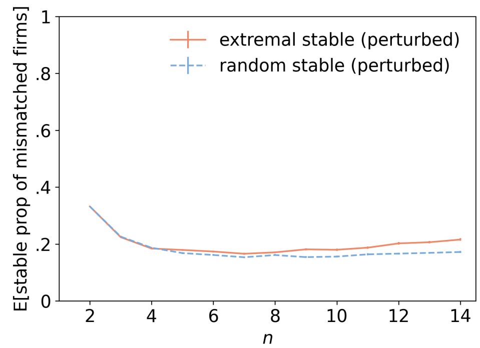
[38] Lebedev et al. (2007) and Ackermann et al. (2011) show the polynomial convergence under uniform dynamics for markets with aligned preferences, a special case of markets with a nested structure of fragments.

[39] Similar results hold for other dynamics. As shown in the Online Appendix, the same insights are obtained for the dynamics where at each step, a randomly-chosen agent—selected uniformly at random among those who have at least one blocking partner—forms a match with her best blocking partner. Analogous results also emerge in various markets with cardinal preferences, where a blocking pair generating a higher total surplus, or a total surplus gain compared to the previous match, is more likely to be formed.

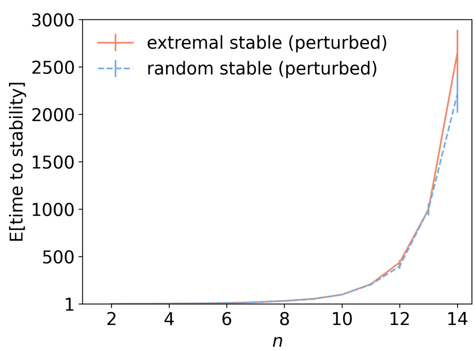
the expected value of these statistics across all markets in the sample.



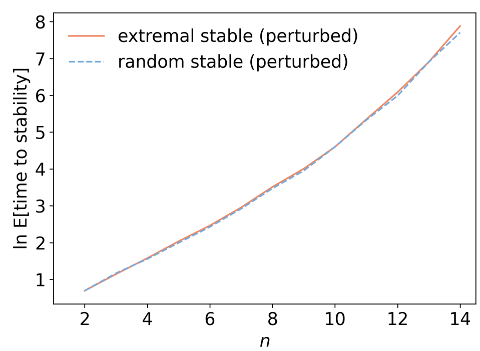
(a) Return probability



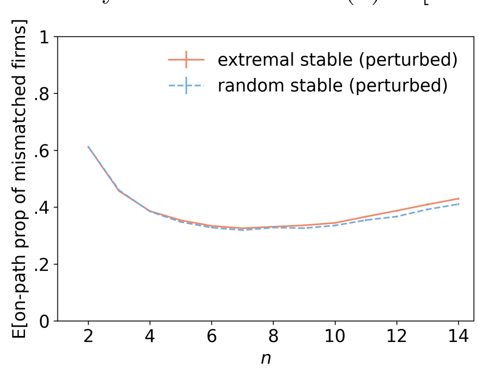
(b) Proportion of mismatched firms/workers in ultimate stable matching



(c) Time to stability



(d) ln [Time to stability]



(e) On-path average proportion of mismatched firms/workers

Figure 2: Random $n \times n$ markets with multiple stable matchings

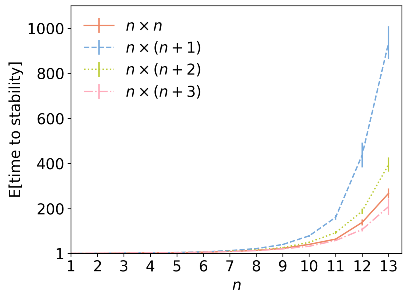
Figure 2 highlights fragility aspects of stable matchings. Even a minimally perturbed stable matching takes an extremely—possibly exponentially—long time to attain a stable matching, not necessarily the perturbed one; see Panels (2c) and (2d) that illustrate the stabilization time and its natural logarithm, respectively. The stabilization process is not only slow but also involves many mismatched agents for long periods, as shown in Panel (2e), which exhibits the on-path average proportion of agents that have different partners compared to the perturbed stable matching. These observations confirm and refine our intuition, discussed in Section 4, regarding the prevalence of "destabilizing" rematching opportunities.

Since decentralized interactions often lead the market away from stability, there is no guarantee whatsoever that the perturbed stable matching returns back or even converges to a nearby, similar stable matching. Panel (2a) shows that indeed the corresponding return probabilities are eventually decreasing, despite a slight non-monotonicity. In markets of size $n = 14$, extremal and random stable matchings, when perturbed minimally, return with a probability of only 40-50%, with the remaining probability converging to other stable matchings. These other stable matchings are considerably different, as demonstrated in Panel (2b), which displays a sizable and non-vanishing proportion of agents whose ultimate stable partners differ from their initial stable partners. Interestingly, extremal stable matchings—a common target of centralized clearinghouses—are more fragile than random stable matchings.[40,41] These findings sharpen the results in Section 3, particularly Corollary 1.

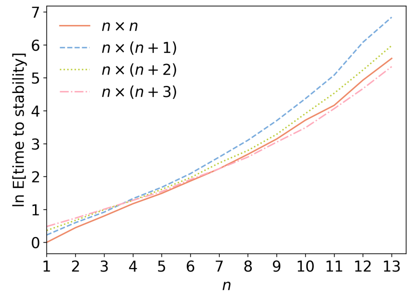
Second, we focus on markets with a unique stable matching. We simulate 1,000 such markets for each market size. For each market, we select one almost stable matching uniformly at random and simulate 300 paths to the stable matching. In addition to balanced $n \times n$ markets, we also consider imbalanced $n \times (n+k)$ markets, both with a fixed imbalance $k \in \{1, 2, 3\}$ and an increasing imbalance $k = n$. This is because for markets with *uniformly random preferences*, imbalanced markets are precisely those markets that have an essentially unique stable matching (Ashlagi, Kanoria, and Leshno, 2017).

[40] The Online Appendix presents two examples that analyze the fragility of stable matchings within a given market. Example 5 provides a market with two completely different stable matchings and shows that one of them is fragile with respect to arbitrary perturbations. Example 6 reports a market in which all stable matchings are fragile and extremal stable matchings are most fragile.

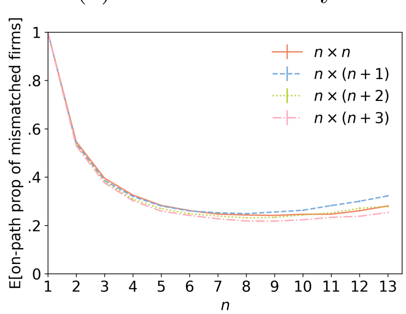
[41] Roth (1989) and Fernandez, Rudov, and Yariv (2022) study outcomes of a centralized clearinghouse in the presence of uncertainty and show that unstable outcomes can emerge in equilibrium.



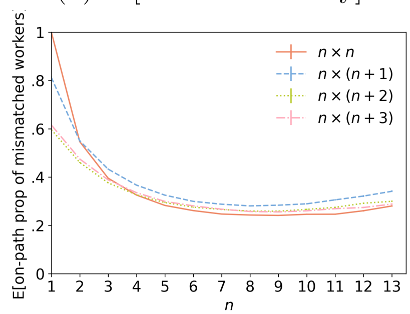
(a) Time to stability



(b) ln [Time to stability]

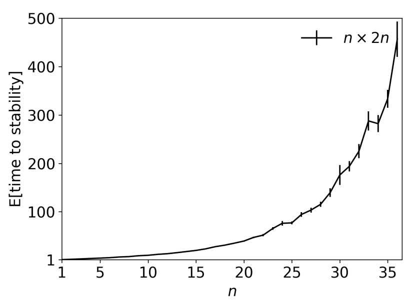


(c) On-path average proportion of mismatched firms

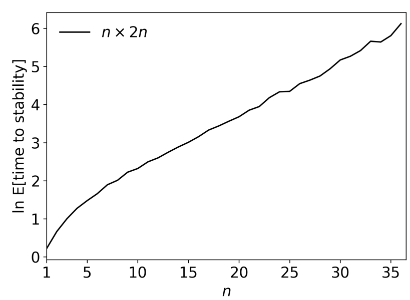


(d) On-path average proportion of mismatched workers

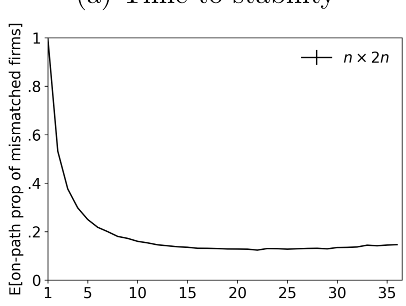
Figure 3: Random $n \times (n + k)$ markets with a unique stable matching, $k \in \{0, 1, 2, 3\}$



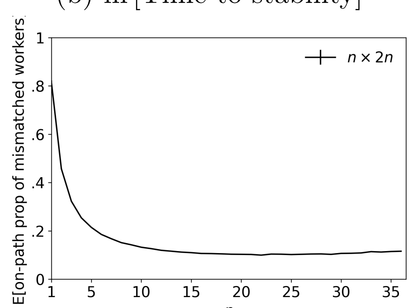
(a) Time to stability



(b) ln [Time to stability]



(c) On-path average proportion of mismatched firms



(d) On-path average proportion of mismatched workers

Figure 4: Random $n \times 2n$ markets with a unique stable matching

Figures 3 and 4 show that unique stable matchings are also fragile, regardless of the size of the imbalance. Although the stabilization process is faster compared to markets with multiple stable matchings, it still lasts for a substantial, and possibly exponentially large, amount of time; see panels (3a), (3b), (4a), and (4b). Furthermore, the dynamics typically entail long periods with a considerable proportion of agents being mismatched, as illustrated by panels (3c), (3d), (4c), and (4d); these panels plot the on-path average proportion of agents whose partners differ from those in their unique stable matching.

6. Discussion

This paper demonstrates the fragility of stable matchings. Even for slight deviations from stability, the stabilization dynamics typically take a very long time, involving many mismatched market participants for extended periods. In markets with multiple stable matchings, stabilization often does not lead back to the slightly perturbed matching or even to a close, similar stable matching. In this sense, stable matchings are not "locally stable." Our results hold for a broad and natural class of dynamics, which provide the decentralized foundation of stable matchings and underlie stability as a solution concept, based on the absence of blocking pairs.[42]

Stable outcomes, whenever they exist, may also be fragile in other environments. For several decades, researchers have sought to identify conditions, sometimes perceived as restrictive, that guarantee the existence of stable outcomes in various settings such as many-to-one and many-to-many markets, roommate markets, and more recently, supply chains and trading networks (e.g., Hatfield et al., 2023), among others; or to find outcomes close to stable when no stable outcome exists.[43] Significant efforts have also been made to establish the decentralized foundation of stable outcomes in these settings, using Roth and Vande Vate (1990)-type dynamics analogous to those in this paper. Our results suggest that even stable outcomes, when they exist, may not be inherently robust.

Our findings highlight the importance of further analysis of the robustness of stable matchings; this area has received much less attention than the robustness analysis of equi-

[42] We show the fragility of stable matchings under favorable conditions for their robustness. We focus on complete-information markets and dynamics that ensure convergence to stability. Furthermore, we suppose that an initial deviation from stability, resulting from some arbitrary shock, is slight; and that there are no additional shocks that could hinder the stabilization process. Despite these favorable conditions, stable matchings are fragile. In settings with incomplete information and changing market conditions, even greater hurdles to stability may emerge.

[43] Little is still known about the core size in many of these settings.

libria, other solution concepts, and mechanisms in economics. It is possible that stable matchings, as currently defined, are less fragile under certain dynamics. The Online Appendix considers dynamics that, at each step, match multiple blocking pairs whenever possible, along with several other reasonable dynamics. However, none of these dynamics even guarantee convergence to stability, which is arguably a minimal requirement for suitable dynamics.[44] Exploring alternative dynamics appears to be a valuable direction for future research. For alternative dynamics, not necessarily relying on formation of blocking pairs, it could be more appropriate to define stability, as a solution concept, more generally, as being based on the absence of beneficial rematching opportunities under these specific dynamics.[45]

There are also other directions for future research. One such direction involves deeper analysis of fragments, a novel concept whose role in various markets is yet to be fully understood, including problems related to algorithmic, probabilistic, and combinatorial aspects. Another direction is using our framework to define numerical robustness measures for stable matchings, which is possible even in ordinal markets, and to compare the fragility levels of different stable matchings within a given market or across markets. Identifying more robust stable matchings may be a promising area.

Appendix A. McVitie and Wilson (1971)'s Algorithm

Gale and Shapley (1962) showed that the set of stable matchings is non-empty for any matching market. In doing so, they introduced the deferred acceptance algorithm (DA) described below:

> Step 1. Each firm proposes to her favorite (acceptable) worker (if any). Each worker who receives more than one offer, tentatively holds on to his favorite (acceptable) offer (if any), and rejects all other offers.
>
> Step k. Each firm who was rejected in step $(k-1)$ makes an offer to her favorite (acceptable) worker who has not rejected her yet (if any). Each worker holds his most preferred acceptable offer to date (if any), and rejects all other offers.
>
> Stop. When no further proposals are made, match each firm to the worker (if any) whose proposal he is holding.

[44] This also applies to the dynamics that, when satisfying a blocking pair, require divorced agents to match with each other (Tamura, 1993; Tan and Su, 1995).

[45] In the school choice context, stability corresponds to elimination of justified envy and is also regarded as a normative fairness criterion (Kojima and Manea, 2010).

The description above corresponds to the firm-proposing DA. Gale and Shapley (1962) proved that its output is the firm-optimal stable matching μ_F. Namely, each firm weakly prefers her assigned partner under this matching to the partner she would get in any other stable matching. The worker-proposing DA defined in a similar way produces the worker-optimal stable matching μ_W. In fact, firms and workers have opposing interests with regard to stable matchings: if firms prefer one stable matching over another, workers hold the opposite opinion (Knuth, 1976). In particular, the firm-optimal stable matching is also worker-pessimal; similarly, the worker-optimal stable matching is firm-pessimal.[46] Furthermore, the set of unmatched agents is the same across all stable matchings, according to the so-called "rural hospital theorem" proved by McVitie and Wilson (1970).

McVitie and Wilson (1971) developed an alternative sequential version of the firm-proposing DA. In their version, proposals by firms to workers are made one at a time and in an arbitrary order, but not simultaneously as in Gale and Shapley (1962)'s algorithm. However, both versions result in the same set of proposals and produce the same matching, the firm-optimal stable matching μ_F.

Importantly, McVitie and Wilson (1971) also adapted the firm-proposing DA, by using an operation called *breakmarriage*, to determine the set of all stable matchings starting from the firm-optimal stable matching. In their paper, they considered a balanced market, i.e., $n = m$, with all agents being acceptable. In what follows, we provide extended results—potentially of independent interest—allowing for a market imbalance and unacceptable agents and connect them to our decentralized dynamics; see Lemma 2 and discussion at the end of this section.[47]

For any stable matching μ and any paired firm f with $\mu(f) = w \in W$, the operation $breakmarriage(\mu, f)$ is defined as follows. It first "breaks" the (f, w)-partnership making firm f free and worker w "semi-free" in the sense that he will only accept a proposal from a firm he prefers to f (Gusfield, 1987); all other agents are initially matched as in μ. Then, the operation forces firm f to propose to the worker following w in her list, thus initiating a sequence of further proposals, rejections, and (tentative) acceptances according to the restarted firm-proposing DA. As there is at most one free firm that has not yet run out of her acceptable alternatives at any time,[48] the corresponding process is deterministic and terminates when (i) worker w receives a proposal from a firm he prefers to his current partner f, (ii) or some previously matched firm runs out of her acceptable alternatives, (iii) or some

[46] In fact, the set of stable matchings forms a distributive lattice (proved by Conway, see Knuth, 1976).

[47] These extended results can be easily used to generalize our main findings to arbitrary markets.

[48] Indeed, all unmatched firms have already run out of their acceptable options in the initial DA instance.

previously unmatched worker receives a proposal from an acceptable firm. In the first case, the termination is called *successful* and leads to a new stable matching $\mu' \neq \mu$ (Theorem 3 in McVitie and Wilson, 1971, also stated as a lemma below).

Lemma (McVitie and Wilson, 1971)**.** *Consider any stable matching μ and any paired firm f with $\mu(f) = w \in W$. If the operation breakmarriage(μ, f) terminates successfully, the resulting matching μ' is stable.*

Proof. The resulting matching μ' is individually rational by definition. Therefore, it is sufficient to prove that there are no firm-worker blocking pairs.

Suppose that some firm f' prefers worker w' to her current match under μ'. Then, w' must have had a proposal from f'. Since the termination is successful, any worker ends up with his best match among those consisting of all firms who have previously proposed to that worker and the outside option of being unmatched. In particular, w' prefers his current match to f'. Therefore, (f', w') does not block μ', as desired. ∎

In the following lemma, we interpret the breakmarriage operation in terms of the decentralized dynamics of our interest. This lemma plays an important role in proving Theorem 1.

Lemma 2. *Under the conditions from the previous lemma, suppose that breakmarriage(μ, f) terminates successfully and generates stable matching μ'. Then, there exists a finite sequence of best blocking pairs that leads from almost stable matching μ_{-f} to μ'.*

Proof. In the course of the breakmarriage operation, there is exactly one free firm—that has not yet run out of her acceptable alternatives—at any time until the successful termination.

At any point of the process, if free f' eventually gets a tentative match with worker w', they form a blocking pair (f', w') for the corresponding matching at that point.

In fact, (f', w') must be the best blocking pair for w'. Indeed, for the sake of contradiction, suppose that instead (f'', w'), with $f'' \neq f'$, is the most preferred blocking pair for w'. Then, by the algorithm, w' must have had a proposal from f''. This contradicts w' agreeing to tentatively match with less desirable f' at the current point of the process. ∎

The theorem below is the main result in McVitie and Wilson (1971), see Theorem 4 in their paper.

Theorem (McVitie and Wilson, 1971)**.** *Any stable matching $\mu \neq \mu_F$ can be obtained from the firm-optimal stable μ_F by successive applications of the breakmarriage operation.*

Proof. Consider any stable matching $\mu \neq \mu_F$. Then, there is a firm f whose partner under μ is not the same as under μ_F. In fact, by the rural hospital theorem (requiring the set of unmatched agents to be the same across all stable matchings), both partners $\mu(f)$ and $\mu_F(f)$ are workers. Apply $breakmarriage(\mu_F, f)$.

We will show that no previously matched firm $f' \in \mu(W)$ can propose to a worse alternative than $\mu(f')$, and thus the process will terminate successfully.[49] Suppose by contradiction that $f' \in \mu(W)$ is the first firm who is rejected by her partner under μ, i.e., worker $\mu(f')$. Then, worker $\mu(f')$ has already received a proposal from a preferred firm f''. We must have $f'' \in \mu(W)$ as well since all firms outside $\mu(W)$ have already run out of their acceptable options in the initial DA instance. By our assumption, f'' has not yet proposed to worker $\mu(f'')$; otherwise, $\mu(f'')$ rejected f'' before $\mu(f')$ rejected f'. Therefore, f'' strictly prefers $\mu(f')$ to $\mu(f'')$. This implies that stable μ is blocked by $(f'', \mu(f'))$, leading to a contradiction.

Therefore, $breakmarriage(\mu_F, f)$ outputs a stable matching in which no previously matched firm has a worse partner than under μ. If the resulting stable matching is not μ, we can iteratively apply the breakmarriage operation until we reach μ after a finite number of applications. Indeed, the breakmarriage operation always makes all firms weakly worse off with some being strictly worse off. ∎

Notably, using the same arguments as in the proof, we can obtain the worker-optimal stable μ_W by successive applications of breakmarriage operations starting from any stable matching, not only from the firm-optimal stable μ_F. Furthermore, by symmetric arguments as in this section, from the worker-optimal stable μ_W, we can get any stable matching $\mu \neq \mu_W$, now by switching to the worker-proposing DA and using symmetrically defined breakmarriage operations with the roles of firms and workers reversed.

In essence, we can extend the above theorem well beyond initial μ_F: by using appropriate versions of DA and breakmarriage operations, any stable matching μ' can be obtained from any stable matching $\mu \neq \mu'$. We use this observation, together with our interpretation of breakmarriage operations in Lemma 2, to prove Theorem 1, one of our main results.

Appendix B. Proof of Lemma 1 and Theorem 1

In this section, we prove Lemma 1 and Theorem 1 stated in Section 3.

[49] First, the condition (ii) of "unsuccessful" termination would trivially fail. Second, if (iii) some previously unmatched worker w'' received a proposal from an acceptable firm f'', $f'' \in \mu(W)$ would prefer w'' to $\mu(f'')$, and thus stable μ would be blocked by (f'', w''), leading to a contradiction.

Proof of Lemma 1. Up to relabeling, suppose that fragment $(\bar{F}, \bar{W})$ of size $k < n$ is induced by matching $\bar{\mu}$ such that $\bar{\mu}(f_i) = w_i$ for all $i \in [k]$.

Assume by contradiction that for some stable matching μ in the original market, we have $\mu(\bar{F}) \neq \bar{W}$. Then, there exists $\bar{w} \in \bar{W}$ with $\mu(\bar{w}) \notin \bar{F}$. Up to relabeling, $\bar{w} = w_1$.

Step 1. By definition of $\bar{\mu}$,

$$\bar{\mu}(w_1) = f_1 \succ_{w_1} \mu(w_1) \notin \bar{F}.$$

By stability of μ,

$$\mu(f_1) \succ_{f_1} w_1 = \bar{\mu}(f_1).$$

By definition of $\bar{\mu}$, $\mu(f_1) \in \bar{W}$. Also, $\mu(f_1) \neq w_1$. Then, up to relabeling, we can assume $\mu(f_1) = w_2$. Therefore,

$$\mu(f_1) = w_2 \succ_{f_1} w_1 = \bar{\mu}(f_1).$$

Step 2. By stability of $\bar{\mu}$,

$$\bar{\mu}(w_2) = f_2 \succ_{w_2} f_1 = \mu(w_2).$$

By stability of μ,

$$\mu(f_2) \succ_{f_2} w_2 = \bar{\mu}(f_2).$$

By definition of $\bar{\mu}$, $\mu(f_2) \in \bar{W}$. In addition, $\mu(f_2) \neq w_1, w_2$.[50] As before, up to relabeling, assume $\mu(f_2) = w_3$.

We can proceed recursively, setting $\mu(f_i) = w_{i+1}$ for $i < k$, until we reach the final step.

Step k. By stability of $\bar{\mu}$,

$$\bar{\mu}(w_k) = f_k \succ_{w_k} f_{k-1} = \mu(w_k).$$

By stability of μ,

$$\mu(f_k) \succ_{f_k} w_k = \bar{\mu}(f_k).$$

By definition of $\bar{\mu}$, $\mu(f_k) \in \bar{W}$. However, by our previous steps, $\mu(f_k) \neq w_1, w_2, \ldots, w_k$, i.e., $\mu(f_k) \notin \bar{W}$. This delivers a contradiction, and the result follows. ∎

The following observation, stated as Lemma 3, is used in our proof of Theorem 1 below.

Lemma 3. *Suppose that there are no non-trivial fragments in the original market. If there is a trivial fragment, then after removing agents forming this fragment from the original market, the remaining market has no non-trivial fragments as well.*

[50] Indeed, since $\mu(f_1) = w_2$, we must have $\mu(f_2) \neq w_2$.

Proof. Suppose, towards a contradiction, that there is a non-trivial fragment in the remaining market. It is straightforward to verify that when merged with the excluded trivial fragment, it extends to a non-trivial fragment in the original market. ∎

Proof of Theorem 1. It suffices to prove $(ii) \Rightarrow (i) \Rightarrow (iii) \Rightarrow (ii)$.

(ii) ⇒ (i): Consider any unstable matching λ. Then, by (ii), it is sufficient to show that λ can attain an almost stable matching. To prove this, we use the recent stabilization result due to Ackermann et al. (2011) from the computer science literature, which strengthens the classical result of Roth and Vande Vate (1990): for any unstable matching, there exists a finite sequence of best blocking pairs that leads to a stable matching.

By this stabilization result, matching λ can reach some stable matching μ. Then, λ can attain an almost stable matching that is one blocking pair away from μ. Indeed, we can terminate the path to μ right before the last step at which the last blocking pair is satisfied.

(i) ⇒ (iii): Suppose, for contradiction, that there is a non-trivial fragment, induced by matching $\bar{\mu}$. Focus on any unstable matching λ that agrees with $\bar{\mu}$ over the fragment. Then, any stable matching that λ can attain must also agree with $\bar{\mu}$ over the fragment.

Since the fragment is non-trivial, there are some stable matchings that disagree with $\bar{\mu}$ over the fragment; and thus they cannot be reached from λ. This yields a contradiction.

(iii) ⇒ (ii): We prove this implication by induction on the size $n = |F| = |W|$ of markets with no non-trivial fragments. For $n = 1$, the result holds immediately. Assume that it holds for all such markets of size at most $(n - 1)$, where $n \geq 2$.

Consider any market of size n, again with no non-trivial fragments. The remaining proof relies on the following two observations.

Observation 1. Consider an arbitrary almost stable matching λ which corresponds to some stable matching μ. We show that λ can reach either any stable matching—as desired—or any other almost stable matching for stable matching μ. A proof of this crucial observation is given at the end of this section.

Observation 2. Lemma 2 in Appendix A establishes the connection between the decentralized dynamics of our interest and breakmarriage operations in McVitie and Wilson (1971)'s algorithm used to find the set of all stable matchings; for the algorithm's description and more details on the connection, see Appendix A. Roughly speaking, any breakmarriage operation, applied to a stable matching, first "divorces" a stable pair of agents and then "restarts" some version of DA. Lemma 2 shows that for all relevant breakmarriage operations, the restarted DA algorithm following breakups can be emulated by our decentralized dynamics.

By combining these two observations, we obtain the desired result. Indeed, consider an arbitrary almost stable matching λ that corresponds to some stable matching μ. Observation 1 allows to attain any almost stable matching—or, equivalently, to "divorce" any stable pair—for μ; assuming we have not shown that λ can reach any stable matching. This, together with Observation 2, implies that the dynamics can replicate any relevant breakmarriage operation for μ and attain a new stable matching μ'. Similar to the proof of $(ii) \Rightarrow (i)$, since λ can attain stable μ', one of its almost stable matchings λ' can also be attained.

Next, we apply the same reasoning for almost stable λ' corresponding to stable μ'. By Observation 1, any almost stable matching for μ' can be attained; assuming we have not demonstrated that λ'—and, consequently, λ—can reach any stable matching. As before, in conjunction with Observation 2, this allows to attain a new stable matching μ''. By proceeding recursively and taking into account our discussion in the last two paragraphs of Appendix A, it follows that the decentralized dynamics can emulate all relevant breakmarriage operations in McVitie and Wilson (1970)'s algorithm. Specifically, λ can attain any stable matching, as desired. Below, we complete the proof by establishing Observation 1.

Proof of Observation 1. Consider an arbitrary almost stable matching λ. Up to relabeling, $\lambda = \mu_{-f_n}$ for stable matching μ, where $\mu(f_i) = w_i$ for any $i \in [n]$. We want to show that μ_{-f_n} can attain either any stable matching or any other almost stable matching μ_{-f_i}, $i < n$.

Step 1. Consider $\bar{F} \equiv \{f_i\}_{i \leq n-1}$ and $\bar{W} \equiv \{w_i\}_{i \leq n-1}$. Suppose first that $\bar{F}$ and $\bar{W}$ form a fragment. This fragment must be trivial due to the absence of non-trivial fragments. From the definition of a trivial fragment, it follows that stable matching μ is the unique stable matching. Then, μ_{-f_n} can easily attain unique stable μ—and therefore any stable matching, as desired—say, by the stabilization result as in the proof of $(ii) \Rightarrow (i)$; or, more directly, by allowing f_n to choose her best blocking pair (f_n, w_n).

Henceforth, suppose that $(\bar{F}, \bar{W})$ is not a fragment. By definition, there is an "inside" agent $a \in \bar{F} \cup \bar{W}$ that prefers some "outside" agent $a' \notin \bar{F} \cup \bar{W}$ to her stable partner $\mu(a)$.

1. If $a \in \bar{F}$, then, up to relabeling, $a = f_{n-1}$. Also, $a' = w_n$ is the only "outside" worker. Thus, $w_n \succ_{f_{n-1}} \mu(f_{n-1}) = w_{n-1}$. By stability of μ, we have $f_n = \mu(w_n) \succ_{w_n} f_{n-1}$.

   Note that μ_{-f_n} can attain the new almost stable matching $\mu_{-f_{n-1}}$ by sequentially satisfying two blocking pairs, (f_{n-1}, w_n) and (f_n, w_n). This path corresponds to agents

f_{n-1} and w_n that successively choose their best blocking pairs, first f_{n-1} and then w_n. The corresponding blocking pairs are indeed the best by stability of μ.

2. Otherwise, if $a \in \bar{W}$, then, up to relabeling, $a = w_{n-1}$. Now, $a' = f_n$ is the only "outside" firm. By symmetric arguments, μ_{-f_n} can attain $\mu_{-f_{n-1}}$ through the path on which first w_{n-1} and then f_n select their best blocking pairs.

In either case, μ_{-f_n} can attain the new almost stable matching $\mu_{-f_{n-1}}$.

Step 2. Redefine subsets $\bar{F} \equiv \{f_i\}_{i \le n-2}$ and $\bar{W} \equiv \{w_i\}_{i \le n-2}$. Suppose first that $\bar{F}$ and $\bar{W}$ form a fragment. Then, it must be trivial. By the definition of a trivial fragment, agents within the fragment must be matched in accordance with μ for any stable matching in the original market. By Lemma 3, the remaining market without those agents has no non-trivial fragments. When "projected" to the remaining market, μ_{-f_n} remains unstable and, by the induction hypothesis, can attain any stable matching in this smaller market. Since all stable matchings in the original market agree on the removed trivial fragment and continue to be stable when "projected" to the remaining market, it implies that matching μ_{-f_n} can attain any stable matching in the original market. Therefore, the desired result holds.

Henceforth, suppose that $(\bar{F}, \bar{W})$ is not a fragment. By definition, there is an "inside" agent $a \in \bar{F} \cup \bar{W}$ that prefers some "outside" agent $a' \notin \bar{F} \cup \bar{W}$ to her stable partner $\mu(a)$.

1. If $a \in \bar{F}$, then, up to relabeling, $a = f_{n-2}$. Since "outside" $a' \in \{w_{n-1}, w_n\}$, we have either $w_n \succ_{f_{n-2}} \mu(f_{n-2}) = w_{n-2}$ or $w_{n-1} \succ_{f_{n-2}} \mu(f_{n-2}) = w_{n-2}$.

   In the former case, as in *Step 1*, μ_{-f_n} can attain the new almost stable matching $\mu_{-f_{n-2}}$ through the path on which first f_{n-2} and then w_n pick their best blocking pairs.

   In the latter case, similar to *Step 1*, $\mu_{-f_{n-1}}$ can reach $\mu_{-f_{n-2}}$ through the path on which first f_{n-2} and then w_{n-1} select their best blocking pairs.

2. If $a \in \bar{W}$, then, up to relabeling, $a = w_{n-2}$. Now, since "outside" $a' \in \{f_{n-1}, f_n\}$, we have either $f_n \succ_{w_{n-2}} \mu(w_{n-2}) = f_{n-2}$ or $f_{n-1} \succ_{w_{n-2}} \mu(w_{n-2}) = f_{n-2}$.

   By symmetric arguments, the new almost stable matching $\mu_{-f_{n-2}}$ can be attained from μ_{-f_n} either directly or indirectly through $\mu_{-f_{n-1}}$.

Consequently, in either case, μ_{-f_n} can attain new $\mu_{-f_{n-2}}$, in addition to $\mu_{-f_{n-1}}$.

We then proceed recursively until we reach the final step, assuming that we have not established the existence of paths from μ_{-f_n} to all stable matchings in the previous steps.

Step $(n-1)$. Redefine subsets $\bar{F} \equiv \{f_1\}$ and $\bar{W} \equiv \{w_1\}$. Suppose first that $\bar{F}$ and $\bar{W}$ form a fragment. Then, it must be trivial. Therefore, μ_{-f_n} can attain any stable matching, as desired, by Lemma 3 and our induction hypothesis, similar to *Steps* $2, 3, \ldots, n-2$.

Henceforth, suppose that $(\bar{F}, \bar{W})$ is not a fragment. Then, as before, μ_{-f_n} can reach the new almost stable matching μ_{-f_1} either directly or indirectly through some—already shown to be attainable in the previous steps—matchings $\mu_{-f_{n-1}}$, $\mu_{-f_{n-2}}, \ldots$, μ_{-f_2}.

To conclude, the above steps guarantee that μ_{-f_n} can attain either any stable matching or any other almost stable matching μ_{-f_i}, $i < n$. This proves Observation 1, and the desired implication $(ii) \Rightarrow (i)$ follows from our previous arguments. ∎

Appendix C. Proof of Theorem 2

In this section, we prove Theorem 2, stated in Section 4, in two steps. First, we establish a result—analogous to the theorem—for a class of *non-augmented markets* with a unique stable matching; the identified class is non-empty, as shown by Lemma 6 in the Online Appendix. Second, we employ markets from this class to augment arbitrary markets and prove the desired theorem.

Proposition 1. *Focus on κ-random (best) dynamics. Fix any $\eta \in (0,1)$, $\epsilon \in (0,1]$, and $\kappa \geq 1$. Consider any sequence of markets of size $n \in \mathbb{N}$ with a unique stable matching such that for any market in the sequence, (i) any worker except one is preferred by at least an η-proportion of firms to their stable partners; (ii) and similarly, any firm except one is preferred by at least an η-proportion of workers to their stable partners. Then, any corresponding sequence of ϵ-unstable matchings with probability $1 - 2^{-\Omega(n)}$ needs $2^{\Omega(n)}$ steps to regain stability.*

Proof. *Note:* The proof below works irrespective of whether we examine "standard" blocking pairs or best blocking pairs. Thereafter, we omit "best" if there is no confusion.

Consider the market of size n in the given sequence and denote its unique stable matching by μ. Suppose that firm $\bar{f}$ and worker $\bar{w}$ are the corresponding "exclusion" agents, see (i) and (ii) in the premise of the proposition. For any matching λ in this market, let

$$\mathcal{S}(\lambda) = |i \in [n] \;\text{ such that }\; \lambda(f_i) = \mu(f_i)|,$$

denote the number of firms (or, equivalently, workers) matched with their stable partners.

Take an arbitrary starting ϵ-unstable matching. To restore stability, at some point of the decentralized process, the market must be arbitrarily close to the stable matching, $\mathcal{S}(\mu) = n$.

In fact, the dynamics need to enter the region $\mathcal{S}(\lambda) \geq (1-\zeta)n$ for some $\zeta \leq \epsilon$, and never leave it until convergence; this must hold for arbitrarily small $\zeta > 0$. Roughly speaking, below we show that for sufficiently small $\zeta > 0$—to be discussed later—"most" matchings in the considered region have substantially more "destabilizing" blocking pairs than "stabilizing" ones. We then connect the dynamics to a random walk that is heavily biased in the "destabilizing" direction. This bias leads to exponentially long paths to stability.

Consider an arbitrary matching λ from the region of interest, i.e., $\mathcal{S}(\lambda) \geq (1-\zeta)n$. For the next few steps, suppose that λ satisfies the following condition $(\star)$: some agent $a \notin \{\bar{f}, \bar{w}\}$ is unmatched. This condition plays an important role; later, we show how to deal with matchings that do not satisfy the condition $(\star)$.

Types of Blocking Pairs. For such a matching λ satisfying $(\star)$, we define two types of blocking pairs. A blocking pair is *destabilizing** if (i) the corresponding new matching λ', obtained from λ, also satisfies $(\star)$; (ii) *and, moreover,* the blocking pair itself is destabilizing, i.e., $\mathcal{S}(\lambda') = \mathcal{S}(\lambda) - 1$. In contrast, a blocking pair is *stabilizing** if (i) the newly obtained matching λ' does not satisfy $(\star)$; (ii) *or* the blocking pair itself is stabilizing, i.e., $\mathcal{S}(\lambda') = \mathcal{S}(\lambda) + 1$.

Note that for any other blocking pair that is neither destabilizing* nor stabilizing*, the newly obtained matching λ' is "similar" to original λ. That is, $\mathcal{S}(\lambda') = \mathcal{S}(\lambda)$ and both new λ' and original λ satisfy $(\star)$. Since we are interested in proving an exponential lower bound for the convergence time, we can essentially "ignore" such blocking pairs that only prolong the process by obtaining "similar" new λ' from original λ.

Number of (De)stabilizing* Blocking Pairs. We show that there are many more destabilizing* blocking pairs than stabilizing* ones, for sufficiently small $\zeta > 0$. Indeed, by $(\star)$, there is an unmatched agent $a \notin \{\bar{f}, \bar{w}\}$ that, by the premise of the proposition, is preferred by at least an η-proportion of agents from the other side to their stable partners. Since $\mathcal{S}(\lambda) \geq (1-\zeta)n$, i.e., at least a $(1-\zeta)$-proportion of agents are already matched with their stable partners, there are at least $(\eta - \zeta)n$ blocking pairs such that $\mathcal{S}(\lambda') = \mathcal{S}(\lambda) - 1$. (*Note:* The same holds true for best blocking pairs.) Among these blocking pairs, all pairs except possibly one pair necessarily unmatch some other $a' \notin \{\bar{f}, \bar{w}\}$. Therefore, there are at least $(\eta - \zeta)n - 1$ destabilizing* blocking pairs.

As concern stabilizing* pairs, there are at most ζn pairs such that $\mathcal{S}(\lambda') = \mathcal{S}(\lambda) + 1$. Indeed, only firms that are not currently matched to their stable partners can be involved in such blocking pairs, and each such firm can take part in at most one such blocking pair.

Furthermore, it is easy to check that there are at most two blocking pairs such that λ' does not satisfy $(\star)$; in that case, either λ' is perfect or only $\bar{f}$ and $\bar{w}$ are unmatched under λ'. To see this, consider the following three cases.

1. First, suppose that, under λ, agents $\bar{f}$ and $\bar{w}$ are either both matched or both unmatched. Then, λ must have a pair of unmatched agents, $f \neq \bar{f}$ and $w \neq \bar{w}$. If λ has several such pairs, λ' must satisfy $(\star)$. If λ has only one such pair, (f, w), then there is at most one blocking pair that makes λ' violate $(\star)$, i.e., the blocking pair (f, w).

2. Second, suppose that $\bar{f}$ is unmatched and $\bar{w}$ is matched in λ. Therefore, λ involves unmatched $w \neq \bar{w}$. If λ has more unmatched pairs than just $(\bar{f}, w)$, then λ' must satisfy $(\star)$. If $(\bar{f}, w)$ is the only unmatched pair for λ, there are at most two blocking pairs that make λ' violate $(\star)$, the blocking pair $(\bar{f}, w)$ that leads to perfect λ' and the blocking pair $(\lambda(\bar{w}), w)$ that generates $(\bar{f}, \bar{w})$ as the only pair of unmatched agents.

3. Third and finally, suppose that $\bar{f}$ is matched and $\bar{w}$ is unmatched in λ. As in the previous case, we can have at most two blocking pairs that make λ' violate $(\star)$.

Thus, there are at most $\zeta n + 2$ stabilizing* blocking pairs.

Likelihood of (De)stabilizing* Blocking Pairs. The above analysis implies that for any unstable matching λ, satisfying $(\star)$ and belonging to the region $\mathcal{S}(\lambda) \geq (1-\zeta)n$, the ratio of the probability to satisfy a destabilizing* blocking pair to the probability to satisfy a stabilizing* blocking pair is bounded from below by

$$\frac{(\eta - \zeta)n - 1}{\kappa(\zeta n + 2)} \geq \frac{\eta - \zeta}{2\kappa\zeta} \to \infty \text{ as } \zeta \to 0.$$

This, in turn, suggests that the probability to satisfy a destabilizing* blocking pair conditional on satisfying a destabilizing* or stabilizing* blocking pair—as already discussed, we essentially "ignore" all other blocking pairs—is bounded from below by

$$p_{d^\star} \equiv \frac{\eta - \zeta}{\eta + (2\kappa - 1)\zeta} \to 1 \text{ as } \zeta \to 0.$$

In what follows, we discuss how to deal with matchings that do not satisfy the condition $(\star)$. We then prove the desired result by establishing a close connection between our decentralized process and a biased random walk.

Reinstatement of the Condition $(\star)$ for Matchings. Consider a matching, say, λ', that violates the condition $(\star)$. Below, we demonstrate that after at most three additional steps—that is, after satisfying at most three blocking pairs in sequence—the desired condition $(\star)$ is restored for the eventually obtained matching; assuming there is no path of length at most three from λ' to the stable matching μ.

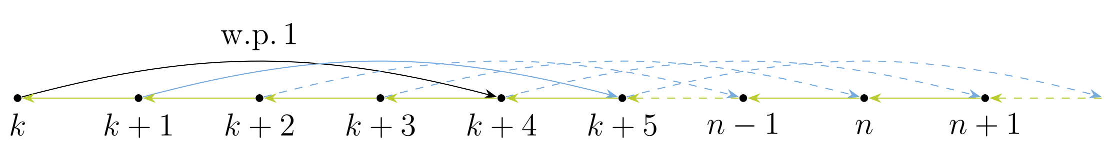
1. If only $\bar{f}$ and $\bar{w}$ are unmatched in the given matching λ', then in the next step, either λ'' satisfies $(\star)$ or λ'' is perfect with $(\bar{f}, \bar{w})$ being matched. In the latter case, after one more step, λ''' must necessarily satisfy the condition $(\star)$. Indeed, there is no way to unmatch only $\bar{f}$ and $\bar{w}$ again.

2. Otherwise, if λ' is perfect, then in the next step, either λ'' satisfies $(\star)$ or only $\bar{f}$ and $\bar{w}$ are unmatched. In the latter case, by using similar arguments as before, we must obtain a matching satisfying $(\star)$ in at most two additional steps.

Link to a Random Walk Heavily Based in the "Destabilizing" Direction. We combine our previous observations to make several relaxations, still allowing us to obtain an exponential lower bound for the convergence time. Let the dynamics start at an unstable matching λ with $\mathcal{S}(\lambda) = \lceil (1-\zeta)n \rceil$, for sufficiently small $\zeta > 0$.

Suppose that the starting matching λ satisfies the condition $(\star)$; if not, this condition must be restored after at most three steps. Then, if the dynamics select to satisfy a destabilizing* blocking pair, the newly obtained matching λ' moves one stable pair away from stability, $\mathcal{S}(\lambda') = \mathcal{S}(\lambda) - 1$, *and* satisfies $(\star)$. In contrast, when a stabilizing* blocking pair is satisfied, new λ' either moves one stable pair towards stability, $\mathcal{S}(\lambda') = \mathcal{S}(\lambda) + 1$, *or* violates the condition $(\star)$. Hereafter, we instead pessimistically assume—based on our analysis above—that after satisfying a stabilizing* blocking pair, new λ' moves four stable pairs towards stability, $\mathcal{S}(\lambda') = \mathcal{S}(\lambda) + 4$, *and satisfies* the condition $(\star)$. This relaxation effectively enables to preserve the condition $(\star)$ throughout the decentralized process.

To obtain the desired lower bound on the number of steps required for our dynamics to converge to the stable matching, $\mathcal{S}(\mu) = n$, we consider the biased random walk $\{S_i\}_{i \geq 0}$



on the set $\{k, k+1, k+2, \ldots\}$, with $k \equiv \lceil (1-\zeta)n \rceil$, defined as follows. This random walk starts at $S_0 = k$. Furthermore, whenever $S_i = k$ for some $i \geq 0$, including starting $i = 0$, the random walk deterministically moves four units to the right, $S_{i+1} = S_i + 4$. Intuitively, if $\mathcal{S}(\lambda)$ ever drops below k for our dynamics, we wait for it to return to that value and restart the random walk. If the "starting" matchings violate the condition $(\star)$, this condition must be restored in at most, say, four steps—in fact, three steps as shown above.

When $S_i > k$ but $S_i < n$—this corresponds to our region of interest in which the condition $(\star)$ is preserved—the random walk moves one unit to the left, $S_{i+1} = S_i - 1$, with probability

$$p_{d^\star} = \frac{\eta - \zeta}{\eta + (2\kappa - 1)\zeta} \to 1 \text{ as } \zeta \to 0,$$

and four units to the right, $S_{i+1} = S_i + 4$, with the remaining probability $p_{s^\star} \equiv 1 - p_{d^\star} \to 0$ as $\zeta \to 0$. Transitions from the remaining nodes $\{n, n+1, n+2, \ldots\}$ are not relevant for our analysis and thus can be defined analogously, for instance.

To conclude, the number of steps required for the random walk $\{S_i\}_{i \geq 0}$ to first reach $S_i \geq n$ is a lower bound for the number of steps it takes for our dynamics to regain stability. The corresponding hitting time for such heavily biased random walks is known to be $2^{\Omega(n)}$ with probability $1 - 2^{-\Omega(n)}$; this result follows from standard Chernoff bound arguments, see Lemma 5 in the Online Appendix. ∎

Proof of Theorem 2. Consider any sequence of *original* markets of size $n \in \mathbb{N}$ and any $\delta > 0$. For the purpose of augmentation, also take some arbitrary sequence of *constructed* markets for which Proposition 1 holds; for instance, the sequence from the proof of Lemma 6 in the Online Appendix. Below, we use these constructed markets to generate the desired δ-*augmented* markets.

Specifically, for every given original market, of size n, add new agents with preferences over each other that replicate the constructed market of size $\lfloor \delta n \rfloor$. In addition, let any agent—original or new—prefer original agents to new agents. It is straightforward to verify that the resulting market, of size $n + \lfloor \delta n \rfloor$, is δ-augmented.

By the construction, the resulting sequence of δ-augmented markets satisfies conditions similar to those in the statement of Proposition 1: there exists $\eta \in (0, 1)$ such that for every market in the sequence, any agent except one "exclusion" firm and one "exclusion" worker is preferred by at least an η-proportion of agents from the other side to their stable partners.

The desired result for this sequence of δ-augmented markets follows from the same arguments as in Proposition 1. Although the market sizes in this sequence are $n + \lfloor \delta n \rfloor$, not n, these sizes still increase linearly with $n \in \mathbb{N}$, so the analogous proof and conclusion hold. ∎

References

Abraham, D. J., P. Biró, and D. F. Manlove (2005): "“Almost Stable” Matchings in the Roommates Problem," in *International Workshop on Approximation and Online Algorithms*, Springer, 1–14.

Accreditation Council for Graduate Medical Education (n.d.): "ACGME Data Resource Book," https://www.acgme.org/about/publications-and-resources/graduate-medical-education-data-resource-book/, accessed July 31, 2024.

Ackermann, H., P. W. Goldberg, V. S. Mirrokni, H. Röglin, and B. Vöcking (2011): "Uncoordinated Two-Sided Matching Markets," *SIAM Journal on Computing*, 40, 92–106.

Agranov, M., A. Dianat, L. Samuelson, and L. Yariv (2023): "Paying to Match: Decentralized Markets with Information Frictions," *mimeo.*

Akbarpour, M., S. Li, and S. O. Gharan (2020): "Thickness and Information in Dynamic Matching Markets," *Journal of Political Economy*, 128, 783–815.

Artemov, G., Y.-K. Che, and Y. He (2023): "Stable Matching with Mistaken Agents," *Journal of Political Economy Microeconomics*, 1, 270–320.

Ashlagi, I., Y. Kanoria, and J. D. Leshno (2017): "Unbalanced Random Matching Markets: The Stark Effect of Competition," *Journal of Political Economy*, 125, 69–98.

Aziz, H., P. Biró, S. Gaspers, R. de Haan, N. Mattei, and B. Rastegari (2020): "Stable Matching with Uncertain Linear Preferences," *Algorithmica*, 82, 1410–1433.

Baccara, M. and L. Yariv (2023): "Dynamic Matching," *Online and Matching-Based Market Design.*

Balinski, M. and G. Ratier (1997): "Of Stable Marriages and Graphs, and Strategy and Polytopes," *SIAM Review*, 39, 575–604.

Biro, P., A. Hassidim, A. Romm, R. I. Shorrer, and S. Sovago (2022): "The Large Core of College Admission Markets: Theory and Evidence," in *Proceedings of the 23rd ACM Conference on Economics and Computation*, 958–959.

Biró, P. and G. Norman (2013): "Analysis of Stochastic Matching Markets," *International Journal of Game Theory*, 42, 1021–1040.

Blum, Y., A. E. Roth, and U. G. Rothblum (1997): "Vacancy Chains and Equilibration in Senior-Level Labor Markets," *Journal of Economic theory*, 76, 362–411.

Boehmer, N., K. Heeger, and R. Niedermeier (2022): "Deepening the (Parameterized) Complexity Analysis of Incremental Stable Matching Problems," *arXiv preprint arXiv:2208.01563.*

Boudreau, J. (2012): "An Exploration into Why Some Matchings Are More Likely than Others," in *The proceedings of MATCH-UP*, 39–50.

Boudreau, J. W. (2011): "A Note on the Efficiency and Fairness of Decentralized Matching," *Operations Research Letters*, 39, 231–233.

Boyd, D., H. Lankford, S. Loeb, and J. Wyckoff (2013): "Analyzing the Determinants of the Matching of Public School Teachers to Jobs: Disentangling the Preferences of Teachers and Employers," *Journal of Labor Economics*, 31, 83–117.

Brilliantova, A. and H. Hosseini (2022): "Fair Stable Matching Meets Correlated Preferences," in *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*.

Chade, H., J. Eeckhout, and L. Smith (2017): "Sorting through Search and Matching Models in Economics," *Journal of Economic Literature*, 55, 493–544.

Chen, Y.-C. and G. Hu (2020): "Learning by Matching," *Theoretical Economics*, 15, 29–56.

Chiappori, P.-A. (2020): "The Theory and Empirics of the Marriage Market," *Annual Review of Economics*, 12, 547–578.

Chiappori, P.-A. and B. Salanié (2016): "The Econometrics of Matching Models," *Journal of Economic Literature*, 54, 832–861.

Chung, K.-S. (2000): "On the Existence of Stable Roommate Matchings," *Games and Economic Behavior*, 33, 206–230.

Clark, S. (2006): "The Uniqueness of Stable Matchings," *Contributions in Theoretical Economics*, 6, 1–28.

Combe, J., U. Dur, O. Tercieux, C. Terrier, and M. Ünver (2022): "Market Design for Distributional Objectives in (Re)assignment: An Application to Improve the Distribution of Teachers in Schools," *mimeo*.

Combe, J., O. Tercieux, and C. Terrier (2022): "The Design of Teacher Assignment: Theory and Evidence," *The Review of Economic Studies*, 89, 3154–3222.

Cwalina, T. B., R. S. Mazmudar, J. S. Bordeaux, and J. F. Scott (2022): "Dermatologist Workforce Mobility: Recent Trends and Characteristics," *JAMA Dermatology*, 158, 323–325.

Diamantoudi, E., E. Miyagawa, and L. Xue (2004): "Random Paths to Stability in the Roommate Problem," *Games and Economic Behavior*, 48, 18–28.

Doğan, B. and L. Ehlers (2021): "Minimally Unstable Pareto Improvements over Deferred Acceptance," *Theoretical Economics*, 16, 1249–1279.

Dworczak, P. (2021): "Deferred Acceptance with Compensation Chains," *Operations Research*, 69, 456–468.

Echenique, F., R. Gonzalez, A. J. Wilson, and L. Yariv (2022): "Top of the Batch: Interviews and the Match," *American Economic Review: Insights*, 4, 223–38.

Echenique, F., A. Robinson-Cortés, and L. Yariv (2023): "An Experimental Study of Decentralized Matching," *mimeo.*

Eeckhout, J. (2000): "On the Uniqueness of Stable Marriage Matchings," *Economics Letters*, 69, 1–8.

Elliott, M. and B. Golub (2022): "Networks and Economic Fragility," *Annual Review of Economics*, 14, 665–696.

Ferdowsian, A., M. Niederle, and L. Yariv (2023): "Decentralized Matching with Aligned Preferences," *mimeo.*

Fernandez, M. A., K. Rudov, and L. Yariv (2022): "Centralized Matching with Incomplete Information," *American Economic Review: Insights*, 4, 18–33.

Fox, J. T. (2009): "Structural Empirical Work Using Matching Models," *New Palgrave Dictionary of Economics. Online edition.*

Gale, D. and L. S. Shapley (1962): "College Admissions and the Stability of Marriage," *The American Mathematical Monthly*, 69, 9–15.

Gangam, R. R., T. Mai, N. Raju, and V. V. Vazirani (2022): "A Structural and Algorithmic Study of Stable Matching Lattices of "Nearby" Instances, with Applications," *arXiv e-prints*, arXiv–1804.

Genc, B., M. Siala, G. Simonin, and B. O'Sullivan (2019): "Complexity Study for the Robust Stable Marriage Problem," *Theoretical Computer Science*, 775, 76–92.

Gul, F. and E. Stacchetti (1999): "Walrasian Equilibrium with Gross Substitutes," *Journal of Economic Theory*, 87, 95–124.

Gusfield, D. (1987): "Three Fast Algorithms for Four Problems in Stable Marriage," *SIAM Journal on Computing*, 16, 111–128.

Haeringer, G. and M. Wooders (2011): "Decentralized Job Matching," *International Journal of Game Theory*, 40, 1–28.

Hall, R. E. and A. B. Krueger (2012): "Evidence on the Incidence of Wage Posting, Wage Bargaining, and On-the-Job Search," *American Economic Journal: Macroeconomics*, 4, 56–67.

Hatfield, J. W., R. Jagadeesan, S. D. Kominers, A. Nichifor, M. Ostrovsky, A. Teytelboym, and A. Westkamp (2023): "Generalized Matching: Contracts and Networks," *Online and Matching-Based Market Design.*

Hatfield, J. W. and S. D. Kominers (2012): "Matching in Networks with Bilateral Contracts," *American Economic Journal: Microeconomics*, 4, 176–208.

Herings, P. J.-J., A. Mauleon, and V. Vannetelbosch (2020): "Matching with Myopic and Farsighted Players," *Journal of Economic Theory*, 190, 105125.

Hoffman, C., A. Levy, and E. Mossel (2023): "Stable Matchings with Correlated Preferences," *arXiv preprint arXiv:2312.14813*.

Hoffman, M., D. Moeller, and R. Paturi (2013): "Jealousy Graphs: Structure and Complexity of Decentralized Stable Matching," in *International Conference on Web and Internet Economics*, Springer, 263–276.

Immorlica, N. and M. Mahdian (2005): "Marriage, Honesty, and Stability," in *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms*, Society for Industrial and Applied Mathematics, 53–62.

Inarra, E., C. Larrea, and E. Molis (2008): "Random Paths to P-Stability in the Roommate Problem," *International Journal of Game Theory*, 36, 461–471.

Jackson, M. O. and A. Watts (2002): "The Evolution of Social and Economic Networks," *Journal of Economic Theory*, 106, 265–295.

Kamada, Y. and F. Kojima (2023): "Fair Matching under Constraints: Theory and Applications," *Review of Economic Studies*, 1–38.

Klaus, B. and F. Klijn (2007): "Paths to Stability for Matching Markets with Couples," *Games and Economic Behavior*, 58, 154–171.

Knuth, D. E. (1976): *Marriages Stables*, Les Presses de l'université de Montréal.

Kojima, F. (2011): "Robust Stability in Matching Markets," *Theoretical Economics*, 6, 257–267.

Kojima, F. and M. Manea (2010): "Axioms for Deferred Acceptance," *Econometrica*, 78, 633–653.

Kojima, F. and P. A. Pathak (2009): "Incentives and Stability in Large Two-Sided Matching Markets," *American Economic Review*, 99, 608–27.

Kojima, F. and M. U. Ünver (2008): "Random Paths to Pairwise Stability in Many-to-Many Matching Problems: a Study on Market Equilibration," *International Journal of Game Theory*, 36, 473–488.

Lazarova, E. and D. Dimitrov (2017): "Paths to Stability in Two-Sided Matching under Uncertainty," *International Journal of Game Theory*, 46, 29–49.

Lebedev, D., F. Mathieu, L. Viennot, A.-T. Gai, J. Reynier, and F. De Montgolfier (2007): "On Using Matching Theory to Understand P2P Network Design," in *INOC 2007, International Network Optimization Conference*.

McVitie, D. G. and L. B. Wilson (1970): "Stable Marriage Assignment for Unequal Sets," *BIT*, 10, 295–309.

——— (1971): "The Stable Marriage Problem," *Communications of the ACM*, 14, 486–490.

Millán, B. and E. Pepa Risma (2018): "Random Path to Stability in a Decentralized Market with Contracts," *Social Choice and Welfare*, 51, 79–103.

Newton, J. and R. Sawa (2015): "A One-Shot Deviation Principle for Stability in Matching Problems," *Journal of Economic Theory*, 157, 1–27.

Pais, J., Á. Pintér, and R. F. Veszteg (2020): "Decentralized Matching Markets with (out) Frictions: a Laboratory Experiment," *Experimental Economics*, 23, 212–239.

Patel, P. N., P. A. Patel, A. H. Sheth, H. Ahmed, T. Begaj, and R. Parikh (2023): "Ophthalmologist Turnover in the United States: Analysis of Workforce Changes from 2014 through 2021," *Ophthalmology*, 130, 973–981.

Pittel, B. (1989): "The Average Number of Stable Matchings," *SIAM Journal on Discrete Mathematics*, 2, 530–549.

Reny, P. J. (2021): "A Simple Sufficient Condition for a Unique and Student-Efficient Stable Matching in the College Admissions Problem," *Economic Theory Bulletin*.

Rheingans-Yoo, R. (2024): "Large Random Matching Markets with Localized Preference Structures Can Exhibit Large Cores," *Games and Economic Behavior*, 144, 71–83.

Roth, A. E. (1989): "Two-Sided Matching with Incomplete Information about Others' Preferences," *Games and Economic Behavior*, 1, 191–209.

Roth, A. E. and E. Peranson (1999): "The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design," *American Economic Review*, 89, 748–780.

Roth, A. E. and M. A. O. Sotomayor (1992): *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*, Econometric Society Monographs, Cambridge: Cambridge University Press.

Roth, A. E. and J. H. Vande Vate (1990): "Random Paths to Stability in Two-Sided Matching," *Econometrica*, 58, 1475–1480.

Rudov, K. (2024): "Decentralized Foundation for Stability of Supply Chain Networks," *mimeo*.

Santavicca, S., D. R. Hughes, H. B. Fleishon, F. Lexa, E. Rubin, A. B. Rosenkrantz, and R. Duszak Jr (2021): "Radiologist-Practice Separation: Recent Trends and Characteristics," *Journal of the American College of Radiology*, 18, 580–589.

Sheth, A. H., V. K. Rathi, G. A. Scangas, L. Xu, M. A. Varvares, and M. R. Naunheim (2023): "Physician Turnover Among Otolaryngologists in the United States, 2014–2021," *The Laryngoscope*, 133, 235.

Shure, K. and Z. Weingarten (2024): "Information Frictions and Teacher Turnover," *mimeo*.

Taie, S. and L. Lewis (2023): "Teacher Attrition and Mobility: Results from the 2021-22 Teacher Follow-Up Survey to the National Teacher and Principal Survey. First Look. NCES 2024-039," *National Center for Education Statistics*.

Tamura, A. (1993): "Transformation from Arbitrary Matchings to Stable Matchings," *Journal of Combinatorial Theory, Series A*, 62, 310–323.

Tan, J. and W. Su (1995): "On the Divorce Digraph of the Stable Marriage Problem," *Proc. National Science Council, Republic of China*, 19, 342–354.

Wang, X., W. Zhang, S. Sheng, R. Sharma, and M. Kamran (2022): "Residency Attrition and Associated Characteristics, a 10-Year Cross Specialty Comparative Study," *Journal of Brain and Neurological Disorders*, 5.