AGGREGATE STABLE MATCHING WITH MONEY BURNING

ALFRED GALICHON[§] AND YU-WEI HSIEH[♣]

Abstract. We propose an alternative notion of non-transferable utility (NTU) stability in matching models that relies on money burning. Our model captures an exchange economy with indivisible goods, fixed prices, and no centralized assignment mechanism. In these models, a non-transferable numéraire (e.g., time) becomes the competitive market-clearing device and enforces the property of equal treatment: two identical individuals will end up with the same equilibrium payoffs. First, we provide a precise connection between our proposed equilibrium concept and the usual NTU stability. Second, by introducing a random utility component, we obtain an NTU counterpart to Choo and Siow's (2006) model. Finally, we provide a dynamic interpretation of the proposed equilibrium concept in a stationary model of market clearing with queues.

Keywords: two-sided matching, non-transferable utility matching, money burning, rationing by waiting, non-price rationing, aggregate matching, matching function, disequilibrium, discrete choice, optimal transport.

JEL Classification: C78, D58

Date: June 12, 2019.

[§] New York University, Department of Economics and Courant Institute of Mathematical Sciences. Address: NYU Economics, 19 W 4th Street, New York, NY 10012. Email: ag133@nyu.edu or galichon@cims.nyu.edu

[♣] University of Southern California, Department of Economics and Dornsife Institute for New Economic Thinking. Address: 3620 South Vermont Ave. Kaprielian Hall Suite 300, Los Angeles, CA 90089. Email: yuwei.hsieh@usc.edu

Acknowledgement: Galichon acknowledges support from NSF grant DMS-1716489, and from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) Grant Agreement no. 295298. This paper has benefited from conversations with Federico Echenique, Christopher Flinn, Jeremy Fox, Douglas Gale, Bryan Graham, Yinghua He, Guy Laroque, Thierry Magnac, Charles Manski, Konrad Menzel, Larry Samuelson, Simon Weber, Glen Weyl, and comments from seminar participants at the Toulouse School of Economics, the Fields Institute for Mathematical Sciences, Carnegie Mellon University, Tepper School of Business, CalTech, UNC, Columbia University, Yale University, Celebrating Chris Flinn's 65th Birthday Conference, and 2015 California Econometrics Conference.

"*Time is Money.*"—Benjamin Franklin—*Advice to a Young Tradesman (1748)*

1. Introduction

The literature on matching markets typically distinguishes between models with transferable utility (TU), in which a numéraire good—often money—clears the market, and models with non-transferable utility (NTU), in which no exchange of a numéraire is permissible. Traditionally, models with transfers have usually been applied to *decentralized* markets such as labor, housing, and marriage markets, whereas models without transfers have typically been used to represent *centralized* markets such as school assignments, organ transplants, and medical residents. In these centralized markets, a market designer clears the market by an algorithmic without prices. Although most markets with transfers are decentralized, and most markets without transfers are centralized, there are important exceptions. The taxi market is a notable example of a *decentralized matching market without transfers*: In this market, the unit fare is fixed by the regulator,[1] and there is no centralized assignment algorithm.

Pairwise stability plays a crucial rule in the centralized NTU matching theory. However, there are two obstacles associated with stability: the problem of aggregation and decentralization. First, it is an equilibrium defined at the *individual* level, making it laborious when the primarily goal is to study the *aggregate* implications. Second, in the absence of a centralized clearinghouse, it is unclear how a stable matching can be achieved. Consider the following motivating example: suppose there are two passengers and one taxi. Passenger 1 and passenger 2 value the ride service by 2 and 1, respectively. The taxi driver is indifferent between the two passengers, and we normalize his utility to zero. Clearly, the *aggregate* outcome is that one passenger gets the ride since there is only one taxi available. However, different matching models have distinct implications about welfare and about *individual* assignments. While assigning the taxi to passenger 1 is one stable matching, assigning it to passenger 2 is a stable matching too. NTU stability itself provides no further guidance about which assignment is more plausible. However, it is not unreasonable to expect that passenger 1 is more likely to obtain the taxi ride: she may be willing to wait longer for the taxi until passenger 2 dropouts—a form of wasteful Bertrand competition. In this story, the

[1] Models without transfers may include the payment of a fee, but the fee should be exogenous to the model.

waiting time decentralizes the efficient allocation by wasteful competition on the demand side. On the other hand, the taxi driver does not benefit from picking up a passenger who has waited for a long time. The social surplus in the model of NTU with waiting is 1. On the other hand, if flexible prices are allowed, the efficient allocation can be decentralized by adjusting the market price to 1, which can be transferred to the driver. The social surplus in the TU set up is 2.[2]

In light of the above example, we propose a model of decentralized NTU matching in which a money-burning mechanism—often time—replaces price as the bidding device. Our model exhibits essential features of both classic TU and NTU matching models. First, it is a *competitive* model akin to TU matching models: waiting lines will form in front of over-demanded agents, and the utility of someone matched with an over-demanded agent will be decreased in proportion to the length of time waited.[3] Second, it is a *non-transferable utility* model: time, unlike price, cannot be transferred from one side of the market to the other, and hence it is pure money burning. We define the concept of "aggregate stable matching with money burning"; such an equilibrium specifies the matching patterns and waiting times associated with each agent. We apply this equilibrium notion to three setups: (1) deterministic utility as in the typical stable matching literature, (2) random-utility discrete-choice models à la Choo and Siow (2006), and (3) a stationary dynamic model with rational expectations. Below we provide an outline for these models and highlight our contributions.

We first investigate the case of deterministic utility and connect our work with Gale and Shapley's (1962) classic theory of stable matching, which is a model based on the individual-level preferences. In the special case when every individual is unique, we show that any stable matching is also an aggregate stable matching supported by no money burning. In

[2] Recently, the surge pricing implemented by a number of ride-sharing companies can be viewed as a wave of moving from the NTU toward the TU mechanism to reduce the waiting time.

[3] To the best of knowledge, rationing-by-waiting can be dated back to the analysis of Barzel (1974). Recently, Margaria (2016) study waiting line from a learning perspective. In a mechanism design context, Braverman et al. (2016) develop a model where the social planner use patients' waiting time to ration over-demanded hospitals. In the context of taxi, however, it is unclear whether it should be taxis waiting for passengers or passengers waiting for taxi, since it mainly depends on the relative number of taxis versus the number of passengers. In contrast to Braverman et al. (2016), in our framework, not only waiting time is endogenous, but also which side of the market shall form the waiting line.

the general case, we can calculate the necessary amount of money burning to decentralize a given stable matching. What makes our analysis novel is that in the decentralized NTU matching models, people do not get their first choices, as one's preferred partner (in the absence of transfers) may prefer to match with another agent. This is part of the reason why the notion of two-sided stability gains popularity, as opposed to the competitive equilibrium in TU matching models. Our solution concept bridges both the TU and NTU approaches. The money burning, such as waiting, plays a similar role to transfers, in the sense that it adjusts supply and demand such that at equilibrium, everyone is *unconditionally* happy with their assignment as in the TU models. In contrast, in the classic notion of stability in NTU matching, in equilibrium, every agent achieves their best option only within the pool of potential partners who rank him or her above their current match.

We next study the case of static, random-utility discrete-choice models, which is known as the "separable" matching models that are largely motivated from empirical perspectives in the wake of Choo and Siow (2006).[4] Recently, separable matching models receive an increasing attention since they can be easily taken to data. To the best of our knowledge, our model is the first separable NTU matching model. We assume agents have (both vertically and horizontally) differentiated utility for being matched to an agent on the other side of the market. We prove the existence and uniqueness of a stable matching with money burning. Interestingly, we show that the equilibrium can be computed by nesting an optimal transport algorithm within a modified version of the Gale-Shapley's deferred acceptance algorithm.[5] Importantly, the NTU matching equilibrium yields a Leontief aggregate matching function, whereas the TU matching equilibrium in Choo and Siow (2006) produces a Cobb-Douglas aggregate matching function.[6] Lastly, we show that the equilibrium in the static random-utility model coincides with the stationary equilibrium in a dynamic model of queueing.

[4] Separable models are models with random heterogeneity in tastes that depend only on the characteristics observed by the econometrician; the preference structure is type-specific, not individual-specific. Examples beyond Choo and Siow (2006) include Galichon and Salanié (2015), and Chiappori, Salanié and Weiss (2017), among others. See a survey in Chiappori and Salanié (2016)

[5] The optimal transport and the deferred acceptance algorithm are fundamental tools for solving classical TU and NTU matching models, respectively.

[6] See Mourifié and Siow (2017) for a survey of the aggregate matching function.

The present paper's contributions are twofold. First, it offers a framework to study decentralized NTU equilibria. This framework leads us to propose a solution concept that complements the classical approach. Our concept is based on an explicit competitive rationing mechanism, and it also permits a natural definition of aggregate stable matching in which an equilibrium exists and is unique. The proposed algorithm to compute equilibrium matching has been implemented.[7]

The rest of the paper is organized as follows. Section 2 considers the deterministic-utility model and compares our solution concept versus the classic stability. Section 3 extends the equilibrium analysis to the random-utility case. Section 4 considers a dynamic model. Section 5 concludes and discuss the role of the proposed model in the literature. To avoid excessive technicality, we offer a brief review of classical discrete-choice theory in section A. In section B, we prove several fundamental results on quantity rationing in discrete choice that will be used to analyze the solution algorithm. These results are of independent interest and can be applied to settings in which goods are available in fixed supply and prices are rigid.[8] The welfare analysis is presented in Appendix C and heavily relies on the tools of convex analysis and optimal transport. Section D summarizes the solution algorithm. All proofs are collected in section E.

2. Aggregate Stable Matching: the Case of Deterministic Utility

2.1. Motivation and Definition.

We illustrate our model by the industry of ride services. We consider the problem of matching different types of cars to different types of passengers. There are $n_x \in \mathbb{N}$ passengers of type $x \in \mathcal{X}$, in which x includes the pick-up location, the size of the party, the type of the ride requested, etc. There are $m_y \in \mathbb{N}$ cars of type $y \in \mathcal{Y}$ available, in which y includes the service offered (e.g., Pool, SUV, or Limo), amenity (e.g., video screen or snack box), and the rating of the driver, etc. Agents are assumed to have preference over types; they are indifferent between two agents with the same type. A type-x passenger enjoys utility α_{xy} from traveling in a car of type y, and a driver of a type-y car enjoys γ_{xy} from serving a passenger of type x. The outside option is labeled by 0, and

[7] In the TraME package (Galichon and O'Hara 2017) http://www.trame-project.com/

[8] The proposed model can potentially be used to infer the demand for golf club membership and the waiting list of prestigious wineries. It is also related to the school choice problem with predetermined quota on gender/ethnic groups.

the corresponding reservation utility of both passengers and drivers is normalized to zero without loss of generality.

We consider the non-transferable utility setup, namely, the price is predetermined—a common practice in the taxi industry. When the market-clearing price is absent, the demand may not be equal to the supply, leading to quantity rationing and waiting line that serve as the market clearing device. We introduce money burning, which is *quasi-linear* in utility: The surplus obtained by a type-x passenger riding in a type-y car after waiting for τ^{α}_{xy} is $\alpha_{xy} - \tau^{\alpha}_{xy}$, whereas the surplus for a type-y car from transporting a passenger of type x after waiting for τ^{γ}_{xy} is $\gamma_{xy} - \tau^{\gamma}_{xy}$. Either passengers or drivers have to wait, depending on which side of the market is in shortage. Formally, in a frictionless market, [9] there cannot exists simultaneously a nonempty waiting line of both passengers and drivers in the market segment xy.

$$\min\left(\tau^{\alpha}_{xy}, \tau^{\gamma}_{xy}\right) = 0. \tag{2.1}$$

We denote by μ_{xy} the number of passengers of type x riding in cars of type y. We consider competitive equilibrium, in which passengers choose the type of cars that maximize their surplus, and taxis choose the type of passengers that maximize their surplus.[10] Let u_x and v_y be the indirect utilities of passengers of type x and taxis of type y, respectively. We have

$$u_x = \max_{y\in\mathcal{Y}}\left\{\alpha_{xy} - \tau^{\alpha}_{xy}, 0\right\} \text{ and } v_y = \max_{x\in\mathcal{X}}\left\{\gamma_{xy} - \tau^{\gamma}_{xy}, 0\right\}. \tag{2.2}$$

Therefore, $u_x - \alpha_{xy} \geq -\tau^{\alpha}_{xy}$ with equality if x chooses y $(\mu_{xy} > 0)$. Similarly, $v_y - \gamma_{xy} \geq -\tau^{\gamma}_{xy}$ with equality if $\mu_{xy} > 0$. As a result,

$$\max\left(u_x - \alpha_{xy}, v_y - \gamma_{xy}\right) \geq \max\left(-\tau^{\alpha}_{xy}, -\tau^{\gamma}_{xy}\right) = -\min\left(\tau^{\alpha}_{xy}, \tau^{\gamma}_{xy}\right) = 0.$$

[9] In the real world, both taxis and passengers may incur non-zero waiting time before the driver reach the pick-up location. Our model abstracts from this friction and only considers the "net" waiting time.

[10] In our model of decentralized NTU matching market, taxis also play an active role in selecting passengers. Traditionally, when passengers book a ride, the dispatch center broadcasts through its network to reach nearby drivers, and the one who response first wins the ride. Today, this practice is largely conducted through the mobile apps. For example, Uber drivers can express their preferences over destinations for two trips everyday. The assignment algorithm will attempt to match them first with passengers who request similar destinations. Our behavior assumption attempts to capture the fact that the drivers still have certain freedom to select the type of passengers *before* boarding, and it does not contradict to the common practice that the taxis cannot refuse to serve according to the destination *after* the consumers are on board.

We conclude that if $\mu_{xy} > 0$, then $u_x - \alpha_{xy} = -\tau^{\alpha}_{xy}$ and $v_y - \gamma_{xy} = -\tau^{\gamma}_{xy}$, and hence $\max\left(u_x - \alpha_{xy}, v_y - \gamma_{xy}\right) = \max\left(-\tau^{\alpha}_{xy}, -\tau^{\gamma}_{xy}\right) = 0$. This brings us to the following definition of an aggregate stable matching:

Definition 1. *An outcome (μ, u, v) is an aggregate stable matching with money burning when the following six conditions are met:*

(i) $\mu_{xy} \in \mathbb{N}$,

(ii) $\sum_{y \in \mathcal{Y}} \mu_{xy} \leq n_x$,

(iii) $\sum_{x \in \mathcal{X}} \mu_{xy} \leq m_y$,

(iv) for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, $\max\left(u_x - \alpha_{xy}, v_y - \gamma_{xy}\right) \geq 0$ *with equality if $\mu_{xy} > 0$,*

(v) for all $x \in \mathcal{X}$, $u_x \geq 0$, with equality if $\mu_{x0} := n_x - \sum_{y \in \mathcal{Y}} \mu_{xy} > 0$, and

(vi) for all $y \in \mathcal{Y}$, $v_y \geq 0$, with equality if $\mu_{0y} := m_y - \sum_{x \in \mathcal{X}} \mu_{xy} > 0$.

Our equilibrium notion is distinct from the classic stable matching, e.g., Gale and Shapley (1962), in two important ways. First, the classic matching theory focuses on deploying a centralized algorithm to achieve a stable matching. In contrast, we introduce a competitive, money-burning mechanism to decentralize the stable matching. In this regard, our approach has a close connection with the transferable utility matching problems studied in Becker (1973) and Shapley-Shubik (1972): by replacing the max function by the summation function in point (iv) of definition 1, one obtains the definition of stable matching with transferable utility. Second, in the classic setup, agents are allowed to express their preference ranking at the individual level. In our setup, by contrast, passengers only care about the *type* of the service. Two cars with distinct license plates but of the same type are perfect substitutes. As a consequence, our notion of stable matching is an *aggregate* equilibrium. Below we elaborate on this point.

2.2. **Comparison with Classic Stable Matching.** Definition 1 is not equivalent to the classic definition of stable matching with non-transferable utility. In this section, we establish the connection between these two equilibrium notions. In particular, the money-burning mechanism provides a way to study the aggregation problem of stable matching. Since Gale and Shapley (1962) is based on describing the matching problem at the individual level, we first need to describe individual passengers and taxis.

We denote by $\mathcal{I}$ the set of passengers and $\mathcal{J}$ the set of taxis. We denote by $x_i \in \mathcal{X}$ the observable type of passenger $i \in \mathcal{I}$ and by $y_j \in \mathcal{Y}$ the observable type of taxi $j \in \mathcal{J}$. If passenger i and taxi j are matched, then the match brings utility α_{ij} to the passenger and γ_{ij} to the taxi. Our assumption implies that individuals in the same observable category are fully indistinguishable; therefore,

$$\alpha_{ij} = \alpha_{x_i y_j}, \gamma_{ij} = \gamma_{x_i y_j}, \alpha_{i0} = \alpha_{x_i 0}, \text{ and } \gamma_{0j} = \gamma_{0 y_j}. \tag{2.3}$$

We augment the choice set by including the outside option $\{0\}$: $\mathcal{I}_0 = \mathcal{I} \cup \{0\}$ and $\mathcal{J}_0 = \mathcal{J} \cup \{0\}$. We further normalize $\alpha_{i0} = 0$ and $\gamma_{0j} = 0$. The matching at the individual level is a binary matrix μ_{ij} such that $\mu_{ij} = 1$ if i and j are matched and 0 otherwise. Moreover, $\mu_{i0} = 1$ ($\mu_{0j} = 1$) if and only if i (j) remains single and 0otherwise. Suppose passenger i and taxi j are matched under μ, their enjoy the utility u_i^μ and v_j^μ, respectively:

$$u_i^\mu = \sum_{j \in \mathcal{J}_0} \mu_{ij} \alpha_{ij} \text{ and } v_j^\mu = \sum_{i \in \mathcal{I}_0} \mu_{ij} \gamma_{ij}. \tag{2.4}$$

Below we summarize the classical definition of stable matching

Definition 2. *The vector μ is a* stable matching in the classical sense *if and only if the following six conditions are met:*

(i) $\mu_{ij} \in \{0, 1\}$,

(ii) $\sum_{j \in \mathcal{J}} \mu_{ij} \leq 1$,

(iii) $\sum_{i \in \mathcal{I}} \mu_{ij} \leq 1$,

(iv) for all $i \in \mathcal{I}$ and $j \in \mathcal{J}$, $\max\left(u_i^\mu - \alpha_{ij}, v_j^\mu - \gamma_{ij}\right) \geq 0$,

(v) for all $i \in \mathcal{I}$, $u_i^\mu \geq 0$, and

(vi) for all $j \in \mathcal{J}$, $v_j^\mu \geq 0$.

Consider a variant of the example in the introduction: suppose that there are two identical passengers and one taxi. The value of being unmatched (for the passengers and the taxi alike) is 0. The value of being matched is 1, both for the passengers and taxi. In the classic model without transfers, there are two individual-level stable matchings, in each of which the matched passenger gets utility one, whereas the unmatched gets utility zero. There is no stable matching in which both passengers get the same payoff. In our model with money burning, we implicitly introduce a third party, say a corrupt taxi dispatcher. The taxi dispatcher allocates the taxi to the passenger who provides the highest bribe which gets accepted, and he refuses the bribe from the unmatched passenger. Clearly, the dominant strategy is for both passengers to provide a bribe worth 1, and the taxi dispatcher then selects randomly one of the bids. There are still two equilibrium matchings *at the individual level*, but in both cases, both passengers get ex-post utility zero, whereas the taxi gets utility 1 (instead of 2 in the case of a transferable utility model). The aggregate stable matching with money burning in this case is $(\mu = 1, u = 0, v = 1)$.

Our first theorem establishes the connection between classic stable matching and the aggregate stable matching with money burning in Definition 1:

Theorem 1. *Assume that* $\sum_{i \in \mathcal{I}} 1\left\{x_i = x\right\} = n_x$ *for all* $x \in \mathcal{X}$ *and* $\sum_{j \in \mathcal{J}} 1\left\{y_j = y\right\} = m_y$ *for all* $y \in \mathcal{Y}$*. Assume that 2.3 holds. Then:*

(i) If μ_{ij} *is a stable matching in the sense of definition 2, then define* u^{μ} *and* v^{μ} *as in (2.4). If one sets*

$$\mu_{xy} = \sum_{i \in \mathcal{I}, j \in \mathcal{J}} \mu_{ij} 1\left\{x_i = x\right\} 1\left\{y_j = y\right\}, \tag{2.5}$$

and

$$u_x = \min_{i : x_i = x} \left\{u_i^{\mu}\right\}, \; v_y = \min_{j : y_j = y} \left\{v_j^{\mu}\right\} \tag{2.6}$$

then (μ, u, v) *is an aggregate stable matching with money buring in the sense of Definition 1.*

(ii) Conversely, assume that (μ, u, v) *is an aggregate stable matching with money burning in the sense of definition 1, then any* μ_{ij} *such that*

$$\mu_{xy} = \sum_{i \in \mathcal{I}, j \in \mathcal{J}} \mu_{ij} 1\left\{x_i = x\right\} 1\left\{y_j = y\right\}$$

is a stable matching in the sense of Definition 2.

The first part of theorem 1 suggests that one may have to burn a given amount of money in order to decentralize a given stable matching in the classic sense. Suppose under the stable matching μ, passenger i of type x is matched with taxi j of type y. One can interpret $\tau_i^{\alpha} = u_i^{\mu} - u_x$ and $\tau_j^{\gamma} = v_j^{\mu} - v_y$ as the times waited. The waiting times are there to ensure that all the agents of the same type receive as much utility as the worse-off agent in that category. The second part of theorem 1 states that any individual-level matching, as long as the aggregate number of matches by types coincides with a given aggregate stable matching with money burning, is also a stable matching. Lastly, theorem 1 implies the following corollary:

Corollary 1. *When there is one individual of each type, any stable matching in the classic sense can be interpreted as an aggregate stable matching supported by no money burning.*

3. Aggregate Stable Matching: The Case of Random Utility

We next study the equilibrium in the random-utility setup. We continue to adopt the language of passengers and taxi drivers as in section 2. There are n_x passengers of type x; a passenger of type x enjoys the systematic utility α_{xy} associated with traveling in a car of type y and an additively separable random-utility components $(\varepsilon_{xy})_y$, whose distribution $\mathbf{P}_x$ may depend on x. Similarly, there are m_y drivers of type y. A driver of a type-y car enjoys the systematic utility γ_{xy} associated with picking up a passenger of type x, and an additively separable random-utility component $(\eta_{xy})_x$, whose distribution $\mathbf{Q}_y$ may depend on y. As in the textbook discrete-choice model reviewed in Appendix A, we assume that each decision maker observes his/her realization of the random-utility component before making the choice. In contrast, the economist who studies the resulting demand system only knows the distributions $(\mathbf{P}_x, \mathbf{Q}_y)$. We make the following assumption on the random-utility component:

Assumption 1. *For all $x \in \mathcal{X}$ ($y \in \mathcal{Y}$), $\mathbf{P}_x$ ($\mathbf{Q}_y$) has a nowhere vanishing density.*

This assumption is valid in many typical models, including the logit case. Our solution concept is a frictionless, competitive-equilibrium analysis à la Choo and Siow (2006), in which agents choose the most preferred type of matching, taking the utility and the amount of money burning as given. The optimal choices from all agents collectively determine the equilibrium money-burning and matching. At equilibrium, (i) demand equals supply, and (ii) there cannot be a pair where money is being burned on both sides of the market, i.e., a passenger of type x waiting for a driver of type y while a driver of type y is simultaneously waiting for a passenger of type x. Formally, we define the aggregate stable matching with money burning as:

Definition 3. *In the case of random utility, an aggregate stable matching with money burning is a vector* $(\mu, \tau^{\alpha}, \tau^{\gamma})$*, where* μ_{xy} *is the number matches of type* xy*,* τ^{α}_{xy} *is the amount of money burned on the side of passengers* x *wishing to match with taxis of type* y*, and* τ^{γ}_{xy} *is the amount of money burned by taxis of type* y *wanting to match with passengers of type* x*, that verify simultaneously:*

- Market Clearing: the number of type-x *passengers who choose a type-*y *ride service under* τ^{α} *equals the number of type-*y *service providers who choose a type-*x *passenger under* τ^{γ}*. Namely, for all* $x \in \mathcal{X}, y \in \mathcal{Y}$*, we have*

$$n_x \Pr\left(y \in \arg\max_{y'} \left\{ \alpha_{xy'} - \tau^{\alpha}_{xy'} + \varepsilon_{xy'} \right\} \right) = m_y \Pr\left(x \in \arg\max_{x'} \left\{ \gamma_{x'y} - \tau^{\gamma}_{x'y} + \eta_{x'y} \right\} \right). \tag{3.1}$$

- One-Sided Money Burning: there is no market xy *where there is a positive money burning both on the passenger and taxi sides; namely,*

$$\min\left(\tau^{\alpha}_{xy}, \tau^{\gamma}_{xy} \right) = 0, \ \forall x \in \mathcal{X}, y \in \mathcal{Y}. \tag{3.2}$$

In the taxi context, Definition 3 implies that the systematic utility is *quasi-linear* in time waited: $U_{xz} = \alpha_{xz} - \tau_{xz}$. Alternatively, one can model time as a discount factor that decreases the utility. However, this paradigm will lead to a non-quasilinear model. It is possible to extend our analysis to such a case by utilizing abstract convex analysis as in Bonnet et al. (2018).

Before we move on to the existence and uniqueness in the general case, the equilibrium defined in Definition 3 admits a closed form expression in the logit case:

Example 1. *If we assume that the random utility terms* $(\varepsilon_{xy})_y$ *and* $(\eta_{xy})_x$ *follow i.i.d. Gumbel distribution, the choice probabilities defined in eq. (3.1) admit a close-form expression, and the resulting system of equations is given by*

$$\begin{aligned}\mu_{xy} = \mu_{x0}\exp\left(\alpha_{xy} - \tau^{\alpha}_{xy}\right) &= \mu_{0y}\exp\left(\gamma_{xy} - \tau^{\gamma}_{xy}\right),\\ \mu_{x0} = n_x/ &\left(1+\sum_y \exp\left(\alpha_{xy}-\tau^{\alpha}_{xy}\right)\right),\\ \mu_{0y} = m_y/ &\left(1+\sum_x \exp\left(\gamma_{xy}-\tau^{\gamma}_{xy}\right)\right).\end{aligned}$$

Therefore,

$$\tau^{\alpha}_{xy} = \alpha_{xy} - \log\left(\mu_{xy}/\mu_{x0}\right), \textit{ and } \tau^{\gamma}_{xy} = \gamma_{xy} - \log\left(\mu_{xy}/\mu_{0y}\right).$$

By imposing condition (3.2), we have

$$\min\left(\alpha_{xy} - \log\left(\mu_{xy}/\mu_{x0}\right), \gamma_{xy} - \log\left(\mu_{xy}/\mu_{0y}\right)\right) = 0,$$

and hence

$$\mu_{xy} = \min\left(\mu_{x0}\exp\left(\alpha_{xy}\right), \mu_{0y}\exp\left(\gamma_{xy}\right)\right). \tag{3.3}$$

By substituting μ_{xy} into the feasibility constraints,

$$\begin{aligned}\mu_{x0} + \textstyle\sum_{y\in\mathcal{Y}}\mu_{xy} &= n_x,\\ \mu_{0y} + \textstyle\sum_{x\in\mathcal{X}}\mu_{xy} &= m_y,\end{aligned} \tag{3.4}$$

one can solve μ_{x0} and μ_{0y} by the following system of equations:

$$\begin{aligned}\mu_{x0} + \textstyle\sum_{y\in\mathcal{Y}}\min\left(\mu_{x0}\exp\left(\alpha_{xy}\right), \mu_{0y}\exp\left(\gamma_{xy}\right)\right) &= n_x,\\ \mu_{0y} + \textstyle\sum_{x\in\mathcal{X}}\min\left(\mu_{x0}\exp\left(\alpha_{xy}\right), \mu_{0y}\exp\left(\gamma_{xy}\right)\right) &= m_y.\end{aligned} \tag{3.5}$$

3.1. **Existence and Uniqueness.** The existence of an unique solution to eq. (3.5) can be established by applying some fixed-point theorem. For the general random taste shifters that go beyond the logit case, however, it is necessarily to introduce a set of mathematical tools since the choice probability does not permit a closed-form expression. For this purpose, in Appendix A, we review the random-utility discrete-choice theory from the perspective of convex analysis and optimal transport. In Appendix B, we further provide several intermediate results that will be used in the subsequent analysis. The equilibrium described in Definition 3 exists under a rather weak condition:

Theorem 2. *Under Assumption 1, the aggregate stable matching with money burning exists.*

The proof is constructive and is based on a ramification of the deferred acceptance algorithm. We describe the algorithm in Appendix D. To show uniqueness, first we define $\tau_{xy} = \tau_{xy}^{\alpha} - \tau_{xy}^{\gamma}$. Clearly, τ_{xy}^{α} and τ_{xy}^{γ} can be treated as the positive part and the negative part of τ_{xy}:

$$\begin{aligned} \tau_{xy}^{\alpha} &= \tau_{xy}^{+} = \max\{\tau_{xy}, 0\}, \\ \tau_{xy}^{\gamma} &= \tau_{xy}^{-} = -\min\{\tau_{xy}, 0\}. \end{aligned} \tag{3.6}$$

Notice that by the definition of positive and negative part of the real-valued function, the condition (3.2) is satisfied automatically. We can characterize the aggregate stable matching with money burning as the solution to a system of nonlinear equations:

$$e\left(\tau\right) = 0, \tag{3.7}$$

where e is the excess-demand function defined by

$$\begin{aligned} e_{xy}\left(\tau\right) \quad &:= \quad n_x \Pr\left(y \in \arg\max_{y'} \left\{\alpha_{xy'} - \tau_{xy'}^{+} + \varepsilon_{xy'}\right\}\right) \\ &- \quad m_y \Pr\left(x \in \arg\max_{x'} \left\{\gamma_{x'y} - \tau_{x'y}^{-} + \eta_{x'y}\right\}\right). \end{aligned} \tag{3.8}$$

The following theorem shows that the solution τ to this equation is unique.[11]

Theorem 3. *Under Assumption 1, the aggregate stable matching with money burning is unique.*

Lastly, we analyze the inefficiency of money burning in Appendix C.

3.2. **Limit when the Stochastic Utility Component is Small.** In this paragraph, we would like to show that the aggregate stable matching with the logit stochastic component studied in section 3 converges (when the amount of randomness tends to zero) to an aggregate stable matching with deterministic utility studied in section 2. To do this, consider a model where the stochastic utility components are logit with scaling parameter $\sigma > 0$. From the analysis in example 1, the equilibrium matching μ_{xy} is given by

$$\mu_{xy}\left(\sigma\right) = \min\left(\mu_{x0}\left(\sigma\right) e^{\alpha_{xy}/\sigma}, \mu_{0y}\left(\sigma\right) e^{\gamma_{xy}/\sigma}\right), \tag{3.9}$$

[11] This result is driven by the fact that the distributions of the random utility components are continuous. By contrast, in the case of deterministic utilities as studied in section 2, there may exist multiple equilibria.

where $\mu_{x0}(\sigma)$ and $\mu_{0y}(\sigma)$ are solution to the system

$$\begin{aligned} &\mu_{x0}(\sigma) + \sum\nolimits_y \min\left(\mu_{x0}(\sigma)\, e^{\alpha_{xy}/\sigma}, \mu_{0y}(\sigma)\, e^{\gamma_{xy}/\sigma}\right) = n_x, \\ &\mu_{0y}(\sigma) + \sum\nolimits_x \min\left(\mu_{x0}(\sigma)\, e^{\alpha_{xy}/\sigma}, \mu_{0y}(\sigma)\, e^{\gamma_{xy}/\sigma}\right) = m_y. \end{aligned} \tag{3.10}$$

Then, the following theorem holds:

Theorem 4. *There are vectors $(u_x) \in \mathbb{R}_+^{\mathcal{X}}$ and $(v_y) \in \mathbb{R}_+^{\mathcal{Y}}$ and a vector $(\mu_{xy}) \in \mathbb{R}_+^{\mathcal{X}\times\mathcal{Y}}$ such that up to subsequence extraction, $u_x = -\lim_{\sigma\to 0} \sigma \ln \mu_{x0}(\sigma)$ and $v_y = -\lim_{\sigma\to 0} \sigma \ln \mu_{0y}(\sigma)$, and (μ, u, v) is the aggregate stable matching with money burning defined in Definition 1.*

4. Interpretation as a stationary dynamic model

We further consider a (discrete-time) dynamic model in which its stationary equilibrium coincides with the equilibrium defined in Definition 3 for static models. The setting here is the same as Section 2 and 3. At each period, there are n_x passengers of type $x \in \mathcal{X}$ appearing in the market, and there are also m_y cars of type $y \in \mathcal{Y}$ joining in the market. Again, the prices are fixed.

The platform tries to clear the market insofar as possible; however, queues must be formed since in general the number of passengers of type x requesting a car of type y at a given time does not coincide with the number of cars of type y opting to pick up a passenger of type x. We let $Q_{xy}^{\alpha}(t)$ be the number of passengers of type x waiting in line to be picked up by a car of type y at time t, and $Q_{xy}^{\gamma}(t)$ be the number of cars of type y queuing to pick up a passenger of type x at time t. The queue is the money burning device that induces waiting times. As in the analysis in Section 2 and 3, we continue to assume utility is quasi-linear in waiting time. Passengers of type x enjoy the gross utility α_{xy} if matched with a type-y car. On the other hand, agents form rational expectations about the waiting time for cars of type y, τ_{xy}^{α}. Due the presence of waiting times, passengers only enjoy the net utility $\alpha_{xy} - \tau_{xy}^{\alpha}$.[12] The passengers can also opt out, in which case their utility is normalized to zero.

For exposition purposes, we assume that there is a logit random taste shifter ε_y in the utility that is observed by the passenger (but not the researchers). As a consequence, the

[12] This additive specification is without loss of generality, as we can measure the utility of passengers in equivalent time units.

proportion of type-x passengers who opt for a car of type y at time t is

$$\frac{\exp\left(\alpha_{xy}-\tau^{\alpha}_{xy}\left(t\right)\right)}{\sum_{y'\in\mathcal{Y}}\exp\left(\alpha_{xy'}-\tau^{\alpha}_{xy'}\left(t\right)\right)+1}.$$

Likewise, drivers of type-y cars enjoy $\gamma_{xy}-\tau^{\gamma}_{xy}$ for picking up a passenger of type x, where γ is the gross utility of drivers, and τ^{γ} is the waiting time. The drivers can drop out, and we normalize their utility to zero. Under the same logit assumption, the proportion of type-y drivers who opt for a passenger of type x at time t is

$$\frac{\exp\left(\gamma_{xy}-\tau^{\gamma}_{xy}\left(t\right)\right)}{\sum_{x'\in\mathcal{X}}\exp\left(\gamma_{x'y}-\tau^{\gamma}_{x'y}\left(t\right)\right)+1}.$$

We shall assume that at all times, the market clears insofar as possible:

$$\min\left(Q^{\alpha}_{xy}\left(t\right),Q^{\gamma}_{xy}\left(t\right)\right)=0. \tag{4.1}$$

Clearly, if there is no queue, there is zero waiting time: $\tau^{\alpha}_{xy}\left(t\right)=0$ if and only if $Q^{\alpha}_{xy}\left(t\right)=0$, and $\tau^{\gamma}_{xy}\left(t\right)=0$ if and only if $Q^{\gamma}_{xy}\left(t\right)=0$.

We further assume that (1) agents are not forward looking; they make decisions solely base on the current waiting time τ, and (2) once the decision is made, they stay in the same queue. The first assumption implies that, at time t, there are

$$Q^{\alpha}_{xy}\left(t\right)+\frac{n_x\exp\left(\alpha_{xy}-\tau^{\alpha}_{xy}\left(t\right)\right)}{\sum_{y'\in\mathcal{Y}}\exp\left(\alpha_{xy'}-\tau^{\alpha}_{xy'}\left(t\right)\right)+1}$$

type x-passengers lining up for type-y cars, the number of those who were queuing at the previous period plus the newly arrived passengers incrementing the queue. Similarly, there are

$$Q^{\gamma}_{xy}\left(t\right)+\frac{m_y\exp\left(\gamma_{xy}-\tau^{\gamma}_{xy}\left(t\right)\right)}{\sum_{x'\in\mathcal{X}}\exp\left(\gamma_{x'y}-\tau^{\gamma}_{x'y}\left(t\right)\right)+1}$$

type y-cars lining up for type x-passengers, again arising from the line at the previous period plus the newly arrived cars.

The total number of xy-matches can possibly be created at time t is therefore the minimum between the number of x-passengers waiting for y-cars, and the number of y-cars waiting for x-passengers, that is

$$\mu_{xy}\left(t\right)=\min\left\{\begin{array}{c} Q_{xy}^{\alpha}\left(t\right)+\frac{n_x\exp\left(\alpha_{xy}-\tau_{xy}^{\alpha}(t)\right)}{\sum_{y'\in\mathcal{Y}}\exp\left(\alpha_{xy'}-\tau_{xy'}^{\alpha}(t)\right)+1},\\ Q_{xy}^{\gamma}\left(t\right)+\frac{m_y\exp\left(\gamma_{xy}-\tau_{xy}^{\gamma}(t)\right)}{\sum_{x'\in\mathcal{X}}\exp\left(\gamma_{x'y}-\tau_{x'y}^{\gamma}(t)\right)+1}\end{array}\right\}, \tag{4.2}$$

and the length of the queues are updated by

$$\left\{\begin{array}{l} Q_{xy}^{\alpha}\left(t+1\right)=Q_{xy}^{\alpha}\left(t\right)+\frac{n_x\exp\left(\alpha_{xy}-\tau_{xy}^{\alpha}(t)\right)}{\sum_{y'\in\mathcal{Y}}\exp\left(\alpha_{xy'}-\tau_{xy'}^{\alpha}(t)\right)+1}-\mu_{xy}\left(t\right)\\ Q_{xy}^{\gamma}\left(t+1\right)=Q_{xy}^{\gamma}\left(t\right)+\frac{m_y\exp\left(\gamma_{xy}-\tau_{xy}^{\gamma}(t)\right)}{\sum_{x'\in\mathcal{X}}\exp\left(\gamma_{x'y}-\tau_{x'y}^{\gamma}(t)\right)+1}-\mu_{xy}\left(t\right)\end{array}\right. \tag{4.3}$$

Clearly, $\min\left(Q_{xy}^{\alpha}\left(t+1\right),Q_{xy}^{\gamma}\left(t+1\right)\right)=0$; therefore

$$\min\left(\tau_{xy}^{\alpha}\left(t+1\right),\tau_{xy}^{\gamma}\left(t+1\right)\right)=0. \tag{4.4}$$

In the stationary state, the lengths of the queues and the waiting times remain constant, and the latter are rationally anticipated by all agents. As a result,

$$\mu_{xy}=\frac{n_x\exp\left(\alpha_{xy}-\tau_{xy}^{\alpha}\right)}{\sum_{y'\in\mathcal{Y}}\exp\left(\alpha_{xy'}-\tau_{xy'}^{\alpha}\right)+1}=\frac{n_x\exp\left(\gamma_{xy}-\tau_{xy}^{\gamma}\right)}{\sum_{x'\in\mathcal{X}}\exp\left(\gamma_{x'y}-\tau_{x'y}^{\gamma}\right)+1} \tag{4.5}$$

and

$$\min\left(\tau_{xy}^{\alpha},\tau_{xy}^{\gamma}\right)=0,$$

which is exactly the aggregate stable matching with money burning for the static model defined in Section 3. It is straightforward to extend to the cases of more general distributions beyond the logit model.

5. Related literature and conclusion

Our paper is related to three streams of the economic literature: (i) non-price rationing, (ii) decentralized matching without transfers, and (iii) matching with unobservable heterogeneity. First, *non-price rationing* arises in many diverse situations such as sticky prices in the macroeconomic theory of disequilibrium, e.g., Bénassy (1976), Gourieroux and Laroque (1985), and Drèze (1987); in credit rationing, e.g., Sealy (1979); in housing market with rent control, e.g., Glaeser and Luttmer (2003); in mechanism design with money burning, e.g., Hartline and Roughgarden (2008) and Braverman et al. (2016); and in health economics, e.g., Lindsay and Feigenbaum (1984), Iversen (1993), Martin and Smith (1999), and Iversen and Siciliani (2011). The mathematical theory of queuing is surveyed in Hassin and Haviv (2003). In econometrics, simultaneous demand/supply systems subject to the quantity rationing constraints have been studied for example by Fair and Jaffee (1972), Gourieroux, Laffont and Monfort (1980), and the survey paper by Maddala (1986). Beyond economics, there is a controversy about the social desirability of waiting lines as a rationing mechanism; a vocal advocate in favor of them is Michael Sandel.[13]

Second, there is a large literature on "market design problems" focused on centralized matching models without transfers, which we will not review here; we shall focus instead on the narrower literature on *decentralized matching without transfers*. Our basic observation is that it is extremely difficult to define the aggregate stable matching when agents are clustered into types of indistinguishable individuals. Indeed, in the absence of transfers, it can be challenging to break ties between identical individuals, [14] and it may therefore be difficult to enforce the desirable requirement that two agents with similar characteristics will obtain the same payoff at equilibrium. Models in the literature have resolved this difficulty mostly by pursuing two approaches. The first approach involves stochastic rationing (see Gale (1996) and references therein) or the introduction of search frictions (see, e.g., Burdett and Coles (1997), Smith (2006) and the references therein). Search frictions provide a way to stochastically ration demand and supply and a rationale to explain variations in the equilibrium payoffs of similar individuals. The second approach involves the introduction of heterogeneity, which can either be observed, as described in a recent paper by Azevedo and Leshno (2016), or unobserved, and can be captured in a random utility model, as described by Dagsvik (2000) and Menzel (2015), who utilize logit heterogeneities. Recently, Che and Koh (2016) have investigated the case of decentralized college admission with uncertain student preferences. In particular, writing college-specific essays can be viewed as a money burning mechanism. See also Echenique and Yariv (2013), and Niederle and Yariv (2009) for other approaches to study decentralized matching markets. Echenique et al. (2013) offer a characterization of rationalizability of matchings without transfers in the spirit of revealed preference.

Third, we consider models with stochastic utility components. Therefore, our paper can be seen as the separable NTU counterpart of the separable TU model with random utility

[13] Consult Sandel (2014).

[14] A literature on fractional stable matchings was initiated with the interesting paper of Roth et al. (1993); however, this model was not designed to handle aggregation problems.

proposed by Galichon and Salanié (2015), who extend the approach of Choo and Siow (2006) beyond the logit case. Galichon et al. (2016) show that by choosing a suitable specification, our model arises naturally as the limiting case of imperfectly transferable utility models with random utility.[15] As described by Azevedo and Leshno (2016), our notion of equilibrium can be interpreted as the solution of a tâtonnement process in a demand and supply framework; however, in contrast to their framework, our framework accommodates a finite number of agents (e.g., the analysis in section 2) and does not require consideration of a continuous limit.

Lastly, it is interesting to contrast our analysis to that of Dagvisk (2000) and Menzel (2015), who study the aggregate implication of NTU stable matching when the number of players is large. First, like in other separable models discussed in Chiappori and Salanié (2016), we require that individual tastes in preferences to be driven by the potential partners' observable type only. This implies that a passenger is indifferent between two cars of the same observable type; in other words, ε_{iy_j} depends on j only through j's observable type y_j. In contrast, the Dagsvik-Menzel model assume independent utility shocks over individual identity, which implies that a passenger is almost never indifferent between two cars, even of the same type. Second, while our analysis allows for any distribution of heterogeneity shocks with that structure, the Dagsvik-Menzel framework intrinsically relies on the logit specification[16].

Interestingly, the Dagsvik-Menzel model yields a Cobb-Douglas matching function with scaling effect, which is given by

$$\mu_{xy} = \mu_{x0}\mu_{0y} \exp(\alpha_{xy} + \gamma_{xy}). \tag{5.1}$$

As shown in example 1, our notion of aggregate stable matching yields the Leontieff matching function:

$$\mu_{xy} = \min\left(\mu_{x0} \exp \alpha_{xy}, \mu_{0y} \exp \gamma_{xy}\right). \tag{5.2}$$

[15] However, Galichon et al. (2016) do not study how the NTU matching can be decentralized and the micro theory of money burning.

[16] Menzel also shows that the model arises as a limit of a number of models that are in the domain of the Gumbel distribution.

On the other hand, Choo and Siow's matching function in the TU separable logit case is given by

$$\mu_{xy} = \sqrt{\mu_{x0}\mu_{0y} \exp(\alpha_{xy} + \gamma_{xy})}. \tag{5.3}$$

Some important remarks are in order.

1. Formula (5.1) resembles Choo-Siow's matching function (5.3). An notable difference is that Choo-Siow's formula is homogeneous of degree zero and hence has no scaling effect (see below). Like in Choo and Siow's case, one cannot separately identify men and women's preference on the basis of formula (5.1), even when the researcher has access to multi-market data.

2. In contrast, the shape of our matching function implies that the identified set of men and women's utilities has "a L" shape, and point identification may be possible when the researcher has multi-market data. Furthermore, parallel to Choo and Siow (2006), the Leontief matching function implied by our model can be estimated from the matching data only, as τ does not appear in the matching function. We plan to investigate the identification and estimation problem in future research.

3. As Menzel (2015) points out, the matching function in (5.1) is not homogeneous. This is because ε_{ij} depend both on the individual man i and the individual woman j, and hence there exists scaling effects: If there are more individual of each types, the expected indirect utility of each individuals is likely to increase to do the benefits of increased diversity.

4. In our setting, our matching function defined by expression (5.2) is homogeneous of degree zero, which is the same as the Choo and Siow's model. This can be understood from the structure of the random utility specification: Because ε_{iy} depends on j only through her observable type y, increasing the number of individuals while maintaining the frequency of the types constant does not affect the equilibrium welfare of the individual participants, as there is no increased benefit of diversity. Therefore, scaling the vector of (n_x, m_y) for all (x, y) by a constant $a > 0$ will simply lead to scaling the number of matches between any pair of types by the same constant, so that μ_{xy} becomes $a\mu_{xy}$.

References

[1] Azevedo, E., and Leshno, J. (2016): A Supply and Demand Framework for Two-Sided Matching Markets. *Journal of Political Economy* 124 (5), pp. 1235–1268.

[2] Barzel, Y. (1974): A Theory of Rationing by Waiting. *Journal of Law and Economics*, 17 (1), pp. 73–95.

[3] Becker, G. S. (1973): A theory of marriage: part I, *Journal of Political Economy*, 81, pp. 813–846.

[4] Braverman, M., J. Chen and S. Kannan (2016): Optimal Provision-After-Wait in Healthcare, *Mathematics of Operations Research*, 41, pp. 352–376.

[5] Bénassy, J.-P. (1976): The Disequilibrium Approach to Monopolistic Price Setting and General Monopolistic Equilibrium, *Review of Economic Studies*, 43, pp 69–81.

[6] Bonnet, O., A. Galichon, K. O'Hara, and M. Shum (2018): Yogurts Choose Consumers? Estimation of Random Utility Models via Two-Sided Matching, working paper.

[7] Burdett, K. and Coles, M. (1997): Marriage and Class, *Quarterly Journal of Economics* 112 (1), pp. 141-168.

[8] Che, Y.-K., and Koh, Y. (2016): "Decentralized College Admissions." *Journal of Political Economy* 124 (5), pp. 1295–1338.

[9] Chiappori, P.-A. and B. Salanié (2016): The Econometrics of Matching Models, *Journal of Economic Literature*, 54(3), pp. 832–861.

[10] Chiappori, P.-A., B. Salanié and Y. Weiss (2017): Partner Choice, Investment in Children, and the Marital College Premium, *American Economic Review*, 107, pp. 2109-2167.

[11] Choo, E., and A. Siow (2006): Who Marries Whom and Why, *Journal of Political Economy*, 114(1), pp. 175–201.

[12] Dagsvik, J. (2000): Aggregation in Matching Markets, *International Economic Review* 41 (1), pp. 27-57.

[13] Drèze, J. (1987): Underemployment Equilibria: From Theory to Econometrics and Policy, *European Economic Review* 31: pp. 9–34.

[14] Echenique, F. and Galichon, A. (2016): Ordinal and Cardinal Solution Concepts for Two-Sided Matching, *Games and Economic Behaviour*, forthcoming.

[15] Echenique, F. Lee, S.M., Shum, M. and Yenmez, M.B. (2013). "The Revealed Preference Theory of Stable and Extremal Stable Matchings." *Econometrica* 81, pp. 153–171.

[16] Echenique, F. and L. Yariv (2013): An Experimental Study of Decentralized Matching, Working Paper.

[17] Fair, R. C. and D. M. Jaffee (1972): Methods of Estimation for Markets in Disequilibrium, *Econometrica*, 40(3), pp. 497–514

[18] Gale, D. and L. S. Shapley (1962): College Admissions and the Stability of Marriage, *The American Mathematical Monthly*, 69(1), pp. 9–15.

[19] Gale, D. (1996): Equilibria and Pareto Optima of Markets with Adverse Selection, *Economic Theory*, 7, pp. 207–235.

[20] Galichon, A. and Salanié, B. (2015): Cupid's invisible hands, working paper.

[21] Galichon, A. and Salanié, B. (2017): The Econometrics and Some Properties of Separable Matching Models, *American Economic Review (Papers and Proceedings)*, 107, 251-255.

[22] Galichon, A. Kominers, S. and Weber, S. (2016): Costly Concessions: An Empirical Framework for Matching with Imperfectly Transferable Utility. *Journal of Political Economy*, forthcoming.

[23] Glaeser, E. and Luttmer, E. (2003). “The Misallocation Of Housing Under Rent Control.” *American Economic Review* 93(4), pp. 1027–1046.

[24] Gourieroux, C., J. J. Laffont and A. Monfort (1980): Disequilibrium Econometrics in Simultaneous Equations Systems, *Econometrica*, 48 (1), pp. 75–96.

[25] Gourieroux, C., and Laroque, G, (1985): The Aggregation of Commodities in Quantity Rationing Models, *International Economic Review* 26 (3), pp. 681–699.

[26] Hartline, J. and T. Roughgarden (2008): Mechanism Design and Money Burning, *STOC*.

[27] Hassin, R. and Haviv, M. (2003): *To Queue or Not to Queue: Equilibrium Behavior in Queuing Systems*, Kluwer Academic Publishers.

[28] Iversen, T. (1993): A Theory of Hospital Waiting List, *Journal of Health Economics*, 12, pp. 55–71.

[29] Iversen, T. and L. Siciliani (2011): Non-Price Rationing and Waiting Times, *Oxford Handbook of Health Economics*, 649–670, Oxford University Press.

[30] Lindsay, C. M. and B. Feigenbaum (1984): Rationing by Waiting Lists, *American Economic Review*, 74(3), pp. 404–417.

[31] Maddala, G. S. (1986): Disequilibrium, Self-Section, and Switching Models, *Handbook of Econometrics*, vol III, pp. 1632–1688.

[32] Margaria, C. (2016). Queuing to learn. Working paper.

[33] Martin, S. and P. C. Smith (1999): Rationing by Waiting Lists: An Empirical Investigation, *Journal of Public Economics*, 71, pp. 141–164.

[34] McFadden, D. (1976): The Mathematical Theory of Demand Models, in *Behavioral Travel-Demand Models*, ed. by P. Stopher, and A. Meyburg, pp. 305–314. Heath and Co.

[35] Menzel, K. (2015): Large Matching Markets as Two-Sided Demand Systems, *Econometrica*, 83(3), pp. 897-941.

[36] Mourifie, I. and A. Siow (2017): The Cobb Douglas Marriage Matching Function: Marriage Matching with Peer and Scale Effects, working paper.

[37] Niederle, M. and L. Yarive (2009): Decentralized Matching with Aligned Preferences, *NBER* Working Paper Number 14840

[38] Rheinboldt, W. (1974): *Methods of solving systems of nonlinear equations*, SIAM.

[39] Rust, J. (1994): Structural estimation of Markov decision processes. *Handbook of Econometrics* IV, chapter 51, pp. 3081–3143.

[40] Rockafellar, R.T. and Wets, R. (2009). *Variational Analysis.* Springer.

[41] Roth, A., Rothblum, U., and Vande Vate, J. (1993): Stable matchings, optimal assignments, and linear programming. *Mathematics of Operations Research.* 18 (4), pp. 803–828.

[42] Roth, A. E. and M. A. O. Sotomayor (1990): *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*, Econometric Society Monographs No. 18, Cambridge University Press.

[43] Sandel, M. (2013). *What Money Can't Buy: The Moral Limits of Markets.* Farrar, Straus and Girous.

[44] Sealy, C. W., Jr. (1979): Credit Rationing in the Commercial Loan Market: Estimates of a Structural Model Under Conditions of Disequilibrium, *Journal of Finance*, 34(2), pp. 689–702.

[45] Shapley, L. S. and M. Shubik (1972): The Assignment Game, I: The Core, *International Journal of Game Theory*, 1, pp. 111–130.

[46] Smith, L. (2006): The marriage model with search frictions. *Journal of Political Economy* 114 (6), pp. 1124–1144.

[47] Topkis, D. (1998): *Supermodularity and Complementarity*, Princeton University Press.

[48] Train, K. E. (2009). *Discrete choice methods with simulation*, Cambridge university press.

Appendix A. Reminders on the Classical Discrete-Choice Theory

To prove the existence and uniqueness results highlighted in section 3, we shall extend the analysis of Galichon and Salanié (2015)—a model with market-clearing prices—to a model without prices. Our analysis heavily relies on the convex analytical properties of the discrete-choice theory for the following three reasons. First, it allows for a general class of random taste shifters. Second, it offers a convenient framework to (1) derive the demand from the utility parameters, (2) derive the inverse demand, and (3) calculate the welfare. Third, the properties of the demand function, such as *gross substitute*, are readily available from the convexity of the welfare function.

In this section, we review two important results in the classical discrete-choice theory: First, the *Daly-Zachary-Williams theorem* that relates the welfare and the demand function. Second, the *entropy of choice* that relates the inverse demand function and the theory of *optimal transport*, which can be deployed for computation. In Appendix B, we show that these two fundamental results can be generalized to the case with exogenous quantity limit. These results forms the foundation for the aggregate stable matching with money burning, in which the quantity limits are endogenously determined by the two-sided, discrete-choice demand system.

A.1. **Demand.** We adopt the classic random-utility framework (e.g., McFadden (1976)), where an agent of type $x \in \mathcal{X}$ is facing a set of choices $z \in \mathcal{Z}_0 := \mathcal{Z} \cup \{0\}$, where 0 is the outside option. We denote by n_x the number of type-x agents. We denote by U_{xz} the systematic utility associated with option $z \in \mathcal{Z}$; the systematic utility associated with the outside option is normalized to zero. In what follows, the vector $U = (U_{xz})$ of the systematic utility indices will be treated as a variable. The consumer has an additive "random" utility term ε_{xz} for alternative $z \in \mathcal{Z}_0$. As in the textbook discrete-choice model, we assume that the decision makers observe ε_{xz}. However, the realization of ε_{zx} is unobservable to the economist; she only knows that it follows the distribution $\mathbf{P}_x$ that may depends on x. Throughout the rest of the paper, we maintain Assumption 1 about $\mathbf{P}_x$.

The (aggregate) demand for alternative $z \in \mathcal{Z}$ from consumers of type x, denoted by $g_{xz}(U)$, is the number of consumers of type x, n_x, times the proportion of type-x consumers who consider alternative z as the best choice:

$$g_{xz}(U) := n_x \mathbf{P}_x \left(U_{xz} + \varepsilon_{xz} \geq \max_{z' \in \mathcal{Z}} \{U_{xz'} + \varepsilon_{xz'}, \varepsilon_0\} \right). \tag{A.1}$$

The definition of demand calls for *integrating* over the distribution of ε. Alternatively, the Daly-Zachary-Williams theorem provides another characterization of demand based on the *derivative*. Consider the following welfare function,[17] defined as the weighted sum of the average indirect utility of different types of consumers:

$$G(U) := \sum_{x \in \mathcal{X}} n_x \mathbb{E}_{\mathbf{P}_x} \left[\max_{z \in \mathcal{Z}} \{U_{xz} + \varepsilon_{xz}, \varepsilon_0\} \right]. \tag{A.2}$$

The Daly-Zachary-Williams theorem says that the type-x consumers' demand for alternative z, $g_{xz}(U)$, can be expressed as the derivative of the welfare function $G(U)$ with respect to the systematic utility U_{xz}:

$$g_{xz}(U) = \partial G(U) / \partial U_{xz}. \tag{A.3}$$

In light of the Daly-Zachary-Williams theorem, one can study the discrete-choice demand by the corresponding welfare function. As will become clear later, while it is difficult to directly express the demand, it is relatively easy to characterize the welfare function in the presence of quantity rationing and two-sided preferences.

[17] The welfare function is often used to measure the changes in consumer surplus in the classical merger analysis; see, e.g., Train (2009). In the textbook case, one does not differentiate between types ($n_x = 1$ and $\mathbf{P}_x = \mathbf{P}$ for all $x \in \mathcal{X}$), and hence $G(U)$ equals $\mathbb{E}_{\mathbf{P}} [\max_{z \in \mathcal{Z}} \{U_z + \varepsilon_z, \varepsilon_0\}]$.

A.2. **Entropy of Choice.** The welfare function and its gradient may be difficult to compute in some cases. However, Galichon and Salanié (2015, hereafter GS) find that the *convex conjugate* of the welfare function, which is known as the *entropy* of the discrete-choice problem, has some desirable properties. Formally, the entropy of choice is defined as the Legendre-Fenchel transform of G. For $(\mu_{xz}) \in \mathbb{R}_+^{\mathcal{X}\times\mathcal{Z}}$ such that $\sum_{z\in\mathcal{Z}} \mu_{xz} \leq n_x$, it is given by

$$G^*(\mu) = \sum_{x\in\mathcal{X}} n_x \max_{(U_{xz})\in\mathbb{R}^{\mathcal{X}\times\mathcal{Z}}} \left\{ \sum_{x\in\mathcal{X},z\in\mathcal{Z}} \mu_{xz}U_{xz} - G(U) \right\}. \tag{A.4}$$

As shown by GS, proposition 2, the negative entropy of choice can be equivalently written as the solution to the following optimal transport problem:

$$-G^*(\mu) = \sum_{x\in\mathcal{X}} n_x \max_{\substack{Z\sim\mu_x \\ \varepsilon\sim\mathbf{P}_x}} \mathbb{E}[\varepsilon_Z], \tag{A.5}$$

where the random variable $Z \in \mathcal{Z}_0$ is the choice from the consumer who's random utility component is ε. Given any parametric specification on $\mathbf{P}_x$, the entropy of choice can be easily computed thanks to the optimal-transport representation. GS further show that the surplus function $G(U)$ has the following characterization involving the entropy of choice:

$$G(U) = \sum_{x\in\mathcal{X},z\in\mathcal{Z}} \mu_{xz}U_{xz} + \sum_{x\in\mathcal{X}} n_x \mathbb{E}_{\mathbf{P}_x}[\varepsilon_Z] = \sum_{x\in\mathcal{X},z\in\mathcal{Z}} \mu_{xz}U_{xz} - G^*(g(U)). \tag{A.6}$$

Clearly, the above characterization of the welfare function has two components: one from the deterministic utility and the other one from the random utility. As we shall see later, a similar decomposition can be obtained under rationing or matching. In some cases, like in the logit case, a closed-form expression exists for G and G^*.

Example 2. *When the random utility components $(\varepsilon_{xz})_{z\in\mathcal{Z}_0}$ are distributed with i.i.d. Gumbel distribution for all x, this model boils down to a standard logit model. In this case, it is well known (see e.g., McFadden 1976) that*

$$G(U) = \sum_{x\in\mathcal{X}} n_x \log\left(1 + \sum_{z\in\mathcal{Z}} \exp(U_{xz})\right), \tag{A.7}$$

and, as shown in GS, G^* also has a closed form:

$$G^*(\mu) = \sum_{x\in\mathcal{X}} \sum_{z\in\mathcal{Z}_0} \mu_{xz} \log\frac{\mu_{xz}}{n_x}, \tag{A.8}$$

where μ_{x0} is implicitly defined by $\mu_{x0} = n_x - \sum_{x \in \mathcal{X}} \mu_{xz}$.

A.3. **Comparative statics.**

Lemma 1. *G is convex and submodular and G^* is convex and supermodular.*

Lemma 1 has several economic interpretations. First, recall that the Daly-Zachary-Williams theorem implies that $\nabla G(U)$ is the demand function. By the convex conjugate property, $\nabla G^*(\mu)$ is therefore the inverse demand function. Second, the submodularity of G is equivalent to the *gross substitute* property: the demand for z weakly decreases when the systematic utility associated with alternative z' decreases. Moreover, $\partial g_{xz}(U) / \partial U_{xz} \geq 0$.

Remark A.1. Under Assumption 1, G and G^* are continuously differentiable, and so the second part of the result, namely the supermodularity of G^*, is equivalent to the fact that the inverse demand function $\nabla G^*(\mu)$ is inverse isotone.

Appendix B. Discrete choice under Availability Constraints

To study the matching equilibrium proposed in section 3, we find it fruitful to first study a simplified model—the model of discrete choice under rationing constraints. In particular, we show how to compute the constrained discrete-choice demand system as well as the associated money burning/waiting time and study their properties.

To introduce the idea, let us assume that there are two alternatives 0 and z_1, and that z_1 is subject to a capacity constraint $\bar{\mu}_1$. The systematic utility associated with z_1 is α_1, and the systematic utility associated with 0 is normalized to zero. There is only one type of agent, whose total mass is 1. Without a capacity constraint, the market shares of 0 and z_1 are respectively $1/(1+\exp\alpha_1)$ and $\exp\alpha_1/(1+\exp\alpha_1)$. Let us now introduce the capacity constraints. If $\exp\alpha_1/(1+\exp\alpha_1) < \bar{\mu}_1$, i.e., the capacity constraint is not binding, then the market shares are the same as before. On the other hand, if $\exp\alpha_1/(1+\exp\alpha_1) > \bar{\mu}_1$, then there will be a stationary waiting time τ_1 associated with alternative z_1, such that $\exp(\alpha_1 - \tau_1) / (1 + \exp(\alpha_1 - \tau_1)) = \bar{\mu}_1$. Therefore the stationary waiting time is $\tau_1 = \alpha + \ln(1/\bar{\mu}_1 - 1)$, which will ensure that the market shares of 0 and z_1 are in this case respectively $1-\bar{\mu}_1$ and $\bar{\mu}_1$. The following paragraph extends this simple logic to the multinomial case with general heterogeneity.

In light of the analysis in Appendix A, we first study the welfare function under rationing and its equivalent representation involving the entropy of choice. Likewise, the demand function can be obtained as the gradient of the welfare function—an extension of the Daly-Zachary-Williams theorem to the cases with rationing. Moreover, the properties of the demand function, in particular the *gross substitute* property, can be deducted from the convexity of the welfare function.

B.1. **Welfare.** Assume that a maximum number $\bar{\mu}_{xz}$ of consumers of type x can obtain alternative $z \in \mathcal{Z}_0$. In this case, consumers will compete for the alternative z by money burning, such as waiting in line. Let τ_{xz} be the amount of time that a consumer of type x needs to wait to obtain z. We assume that the consumer's systematic utility associated with alternative z is *quasi-linear* in time waited: $U_{xz} = \alpha_{xz} - \tau_{xz}$. Let μ_{xz} be the number of consumers of type x choosing alternative z. Here τ_{xz} is the Lagrange multiplier of the scarcity constraint $\mu_{xz} \leq \bar{\mu}_{xz}$, and thus τ and μ are determined by the complementary slackness conditions:

$$\begin{aligned} &\mu_{xz} = \partial G\left(\alpha_{xz} - \tau_{xz}\right)/\partial U_{xz}, \\ &\mu_{xz} \leq \bar{\mu}_{xz}, \tau_{xz} \geq 0, \\ &\tau_{xz} > 0 \Longrightarrow \mu_{xz} = \bar{\mu}_{xz}. \end{aligned} \tag{B.1}$$

As it turns out, the vectors μ and τ as the solution to the system (B.1) can be expressed as the gradient with respect to α and $\bar{\mu}$ of the map $(\alpha, \bar{\mu}) \to \bar{G}(\alpha, \bar{\mu})$ defined by

$$\bar{G}\left(\alpha, \bar{\mu}\right) = \min_{\tau \geq 0} \left\{ G\left(\alpha - \tau\right) + \sum_{x \in \mathcal{X}, z \in \mathcal{Z}} \bar{\mu}_{xz} \tau_{xz} \right\}. \tag{B.2}$$

This quantity has an interesting interpretation in terms of welfare analysis. Indeed, at the optimal value of τ,

$$\bar{G}\left(\alpha, \bar{\mu}\right) = G\left(\alpha - \tau\right) + \sum_{x \in \mathcal{X}, z \in \mathcal{Z}} \bar{\mu}_{xz} \tau_{xz}. \tag{B.3}$$

Clearly, $\bar{G}(\alpha, \bar{\mu})$ can be interpreted as the first best, the maximum welfare attainable by a central planner. However, this welfare cannot be attained in a decentralized market; in such a market, one needs a price vector τ to clear demand and supply, and the second best welfare actually achieved by consumers is only $G(\alpha-\tau)$. Consequently, $\sum_{x\in\mathcal{X},z\in\mathcal{Z}}\bar{\mu}_{xz}\tau_{xz}=\sum_{x\in\mathcal{X},z\in\mathcal{Z}}\mu_{xz}\tau_{xz}$ is the total efficiency loss, which is the total time wasted in line—a natural measure of departure from efficiency. For this reason we shall call $\bar{G}$ the *capacity-constrained welfare function*. Parallel to the analysis in Appendix A, $\bar{G}$ admits a representation involving $G^*(\mu)$ defined in (A.4) above.

Proposition 1. *The value of $\bar{G}$ is given by*

$$\begin{aligned}\bar{G}(\alpha,\bar{\mu})=\max_{\mu\geq 0} &\quad \left\{\sum_{x\in\mathcal{X},z\in\mathcal{Z}}\mu_{xz}\alpha_{xz}-G^*(\mu)\right\}. \\ s.t. &\quad \mu_{xz}\leq\bar{\mu}_{xz},\ x\in\mathcal{X},\ z\in\mathcal{Z}\end{aligned} \tag{B.4}$$

Expressions (B.4) and (B.2), which are Legendre-Fenchel transformations, immediately imply the following consequence:

Corollary 2. *$\bar{G}(\alpha,\bar{\mu})$ is convex in α and concave in $\bar{\mu}$.*

We further show that the capacity-constrained welfare function admits an optimal-transport representation. Assume a benevolent social planner were in charge of assigning each individual, characterized by their full type (x,ε), to an alternative $z\in\mathcal{Z}$. Then the social planner's problem amounts to picking a conditional probability $\pi(z|x,\varepsilon)$ of assigning an individual of type (x,ε) to an alternative z. Letting $d\pi(\varepsilon,z|x)=\pi(z|x,\varepsilon)\,d\mathbf{P}_x$ be the induced joint distribution on (ε,z) conditional on the type x, the social planner's problem is to pick $\pi(\varepsilon,z|x)$ so to maximize the overall utility $\sum_{x\in\mathcal{X}}n_x\int(\alpha_{xz}+\varepsilon_z)\,d\pi(\varepsilon,z|x)$ subject to the constraint that under $\pi(\varepsilon,z|x)$, the distribution of ε is $\mathbf{P}_x$, and the probability number of z is less or equal than $\bar{\mu}_{xz}$. Letting $\overline{\mathcal{M}}_x(\mathbf{P},\bar{\mu})$ be the set of such distributions, formally defined as

$$\overline{\mathcal{M}}_x(\mathbf{P},\bar{\mu})=\left\{\pi(\varepsilon,z|x):\sum_z\pi(\varepsilon,z|x)=\mathbf{P}_x(d\varepsilon)\,n_x \text{ and } \int\pi(\varepsilon,z|x)\,d\varepsilon\leq\bar{\mu}_{xz}\right\}. \tag{B.5}$$

Our next result shows that $\bar{G}$ can also be interpreted as the value of the social planner's objective function:

Proposition 2. *The value of $\bar{G}$ can be expressed as the weighted sum of the value of optimal transport problems*

$$\bar{G}\left(\alpha,\bar{\mu}\right)=\sum_{x\in\mathcal{X}}n_{x}\max_{\pi\left(.,.|x\right)\in\overline{\mathcal{M}}_{x}\left(\mathbf{P},\bar{\mu}\right)}\int\left(\alpha_{xz}+\varepsilon_{z}\right)d\pi\left(\varepsilon,z|x\right). \tag{B.6}$$

The computation of $\bar{G}$ therefore amounts to solving $|\mathcal{X}|$ subproblems that are themselves optimal transport problems. This fact is very useful in the numerical applications, as it allows for efficient computation of $\bar{G}$ and its gradient.

B.2. **Demand.** Let us define $\bar{g}\left(\alpha,\bar{\mu}\right)$ and $T\left(\alpha,\bar{\mu}\right)$ as the vectors μ and τ that solve (B.1). It follows immediately from the previous paragraph that the vector number of consumers of each type choosing an option of each type is given by

$$\begin{aligned}\bar{g}\left(\alpha,\bar{\mu}\right)=\operatorname{argmax}_{\mu\geq0}\quad&\left\{\sum_{x\in\mathcal{X},z\in\mathcal{Z}}\mu_{xz}\alpha_{xz}-G^{*}\left(\mu\right)\right\},\\ s.t.\quad&\mu_{xz}\leq\bar{\mu}_{xz},\ x\in\mathcal{X},\ z\in\mathcal{Z}\end{aligned} \tag{B.7}$$

and the vector of waiting times waited in each segment of the market is expressed as

$$T\left(\alpha,\bar{\mu}\right)=\operatorname{argmin}_{\tau\geq0}\left\{G\left(\alpha-\tau\right)+\sum_{x\in\mathcal{X},z\in\mathcal{Z}}\bar{\mu}_{xz}\tau_{xz}\right\}. \tag{B.8}$$

The fact that these maximizers exist and are unique is a consequence of the following result that extends the Daly-Zachary-Williams theorem to the case with rationing:

Proposition 3. *(Constrained Daly-Zachary-Williams Theorem) Under Assumption 1, $\bar{G}$ is continuously differentiable, and one has*

$$\bar{g}_{xz}\left(\alpha,\bar{\mu}\right)=\frac{\partial\bar{G}\left(\alpha,\bar{\mu}\right)}{\partial\alpha_{xz}}\text{ and }T_{xz}\left(\alpha,\bar{\mu}\right)=\frac{\partial\bar{G}\left(\alpha,\bar{\mu}\right)}{\partial\bar{\mu}_{xz}}. \tag{B.9}$$

Combining (B.9) with the first order conditions in (B.8), we see that the maps $\bar{g}$ and T satisfy the property

$$\nabla G\left(\alpha-T\left(\alpha,\bar{\mu}\right)\right)=\bar{g}_{xz}\left(\alpha,\bar{\mu}\right),$$

which means that the unconstrained demand associated to $\alpha-T\left(\alpha,\bar{\mu}\right)$ coincides with the constrained demand in which the systematic utility is α and the capacity constraint is $\bar{\mu}$.

Example 3. *In the logit case, system (B.1) boils down to*

$$\begin{cases} \alpha_{xz} - \tau_{xz} = \log \frac{\mu_{xz}}{\mu_{x0}}, \\ \mu_{xz} \leq \bar{\mu}_{xz}, \tau_{xz} \geq 0, \\ \tau_{xz} > 0 \Longrightarrow \mu_{xz} = \bar{\mu}_{xz}. \end{cases} \tag{B.10}$$

Hence, it follows from the first equation that $\mu_{xz} = \mu_{x0} \exp(\alpha_{xz} - \tau_{xz})$*, and*

$$\mu_{xz} = \min\left(\bar{\mu}_{xz}, \mu^*_{x0} \exp(\alpha_{xz})\right), \tag{B.11}$$

where μ^*_{x0} *is the solution to the scalar equation*

$$\mu^*_{x0} + \sum_{z \in \mathcal{Z}} \min\left(\bar{\mu}_{xz}, \mu^*_{x0} e^{\alpha_{xz}}\right) = n_x. \tag{B.12}$$

This equation has a unique solution given the fact that the left-hand side is a continuous and increasing from $\mathbb{R}_+$ *to* $\mathbb{R}_+$*. In this case,*

$$\bar{g}_{xz}(\alpha, \bar{\mu}) = \min\left(\bar{\mu}_{xz}, \mu^*_{x0} e^{\alpha_{xz}}\right), \text{ and } T_{xz}(\alpha, \bar{\mu}) = \max\left(\alpha_{xy} + \log\left(\mu^*_{x0}/\bar{\mu}_{xz}\right), 0\right). \tag{B.13}$$

B.3. **Comparative statics.** We investigate what happens to the waiting times and to the demand when the availability constraint is tightened (namely, when all the entries of the capacity vector $\bar{\mu}$ weakly decrease). Our first result expresses that when the constraint becomes tighter ($\bar{\mu}$ weakly decreases componentwise), all of the entries of the vector τ weakly increase in the componentwise order. If there was only one market segment, the result would be straightforward: when the capacity decreases, the price (here, the waiting time) increases. But when there are multiple markets segments xz, it is no longer obvious that it should be the case. The fact that the result holds, that is, when one entry of the capacity vector decreases, all the waiting times weakly increase, is not a trivial result and essentially follows from the fact that alternative z are gross substitutes, meaning that a decrease in the availability of one alternative will not lead to a decrease in the exogenous utility associated to another one.

Theorem 5. *Under Assumption 1, the shadow price* $T(\alpha, \bar{\mu})$ *is an antitone function of the vector of number of available offers* $\bar{\mu}$*.*

Theorem 5 generalizes the analysis in Lindsay and Feigenbaum (1984), who show that more doctors can reduce the waiting time of surgery. However, here, we are in the vector case: if *one* entry of the capacity vector $\bar{\mu}_{xy}$ increases, then all of the waiting times τ_{xz} are weakly decreased. This is a consequence of the gross substitute property: if the capacity constrained $\bar{\mu}_{xy}$ associated to the xy segment of the market is loosened, then the corresponding waiting time τ_{xy} is decreased; but the other market segments xy' are subsitute for xy, and thus become less congested, which translated into a decreased waiting time $\tau_{xy'}$ for them too.

Our second result expresses the fact that when the capacity constraints are weakly tightened (namely, when the entries of $\bar{\mu}$ weakly decrease), the number of non-demanded options also weakly decreases in each market segment.

Theorem 6. *Under Assumption 1, the number of nondemanded options $\bar{\mu} - \bar{g}\left(\alpha, \bar{\mu}\right)$, is an isotone function of the capacity vector $\bar{\mu}$.*

The intuition behind theorem 6 is that if the capacity on segment xz is increased, the choices that were dominated are still dominated. This is also a characteristic of the gross substitute property in our model: adding options of type xz does not make another option more attractive. The proof of the result is based on two lemmas.

Lemma 2. *Under Assumption 1, one has*

$$\frac{\partial \bar{g}_{xz}}{\partial \bar{\mu}_{x'z'}} = \frac{\partial T_{x'z'}}{\partial \alpha_{xz}}. \tag{B.14}$$

In the logit case, lemma 2 is illustrated as follows.

Example 2 (continued). *In the logit case, $\partial \bar{g}_{xz}\left(\alpha, \bar{\mu}\right) / \partial \bar{\mu}_{xz} = 1\left\{\bar{\mu}_{xz} \leq \mu^{*}_{x0} e^{\alpha_{xz}}\right\}$ and $\partial T\left(\alpha, \bar{\mu}\right) / \partial \alpha_{xz} = 1\left\{\alpha_{xy} + \log\left(\mu^{*}_{x0} / \bar{\mu}_{xz}\right) \geq 0\right\}$, which obviously coincides with each other.*

Lemma 3. *Under Assumption 1, $\alpha - T\left(\alpha, \bar{\mu}\right)$ is an isotone function of α.*

Appendix C. Welfare Analysis of Money Burning in Matching

We next evaluate the welfare implication of money burning in the matching context, which we call *congestion inefficiency*. Let $l^{G}_{xy}\left(\tau^{\alpha}_{xy}\right)$ be the corresponding social loss if a type-x passenger waits for a taxi of type y an amount of time $\tau^{\alpha}_{xy} \geq 0$. Similarly, let us denote $l^{H}_{xy}\left(\tau^{\gamma}_{xy}\right)$ the social loss for the taxi. The loss functions l_{xy} should satisfy $l_{xy}(0) = 0$ and that $l_{xy}(t) > 0$ for $t > 0$.

If $(\mu, \tau^{\alpha}, \tau^{\gamma})$ is the aggregate stable matching with money burning corresponding to the systematic utility indices (α, γ) and marginal distributions of types n and m, then the total welfare loss is

$$L\left(\alpha, \gamma, n, m\right) = \sum_{xy} \mu_{xy}\left(l^{G}_{xy}\left(\tau^{\alpha}_{xy}\right) + l^{H}_{xy}\left(\tau^{\gamma}_{xy}\right)\right). \tag{C.1}$$

We further define G and H as the welfare functions of passengers and taxis with respect to the systematic utility U and V, respectively:

$$G\left(U\right) = \sum_{x\in\mathcal{X}} n_x \mathbb{E}_{\mathbf{P}_x}\left[\max\left\{U_{xy} + \varepsilon_{xy}, \varepsilon_0\right\}\right] \text{ and } H\left(V\right) = \sum_{y\in\mathcal{Y}} m_y \mathbb{E}_{\mathbf{Q}_y}\left[\max\left\{V_{xy} + \eta_{xy}, \eta_0\right\}\right].$$

Their corresponding convex conjugates, G^* and H^*, are defined as:

$$G^*\left(\mu\right) = \max_{(U_{xy})}\left\{\sum_{x\in\mathcal{X},y\in\mathcal{Y}} \mu_{xy}U_{xy} - G\left(U\right)\right\} \text{ and } H^*\left(\mu\right) = \max_{(V_{xy})}\left\{\sum_{x\in\mathcal{X},y\in\mathcal{Y}} \mu_{xy}V_{xy} - H\left(V\right)\right\}.$$

Our next result shows that the total welfare loss is zero if and only if the market is equivalent to a market with transfers. Equivalently, non-transferable utility matching must lead to an efficiency loss.

Proposition 4. *Assume that the distributions $\mathbf{P}_x$ and $\mathbf{Q}_y$ have a non-vanishing density. Let $(\tau^{\alpha}, \tau^{\gamma})$ be the equilibrium matching under rationing-by-waiting associated with matching distribution μ. The total welfare loss $L\left(\alpha, \gamma, n, m\right)$ is zero if and only if the matching coincides a matching with transferable utility, i.e. if and only if μ maximizes*

$$\max_{\mu\geq 0}\left\{\sum_{xy} \mu_{xy}\left(\alpha_{xy} + \gamma_{xy}\right) - \mathcal{E}\left(\mu\right)\right\} \tag{C.2}$$

where $\mathcal{E}\left(\mu\right) = G^*\left(\mu\right) + H^*\left(\mu\right)$ if $\sum_y \mu_{xy} \leq n_x$ and $\sum_x \mu_{xy} \leq m_y$, $\mathcal{E}\left(\mu\right) = +\infty$ else.

Formulation (C.2), introduced in Galichon and Salanié (2015), characterizes matching models with transferable utility. We note that when $L\left(\alpha, \gamma, n, m\right) = 0$, the equivalent matching market with transferable utility is a market in which there *could* be some transfer, but in which there is none at equilibrium. This is therefore a "no-trade equilibrium" in the sense of Echenique and Galichon (2016).

Appendix D. Random-Utility Deferred Acceptance Algorithm

Let $\mu_{xy}^{A,k}$ be the number of offers that can be made by taxis of type x to passengers of type y at the beginning of step $k+1$. This number should be set high enough so that the number of available offers is not binding at the initial step of the algorithm; hence $\mu_{xy}^{A,0} = n_x$. Let $\mu_{xy}^{P,k}$ be the number of proposals made by taxis of type x to passengers of type y at step k. This number should arise from the maximization of taxis' utility under their availability constraint; hence, $\mu_{xy}^{P,k} = \bar{g}_{xy}\left(\alpha, \mu^{A,k-1}\right)$. Let $\mu_{xy}^{T,k}$ be the number of offers from taxis of type x that are tentatively accepted by passengers of type y. Passengers of type y maximize their utility among the proposals that were made to them at step k; hence $\mu_{xy}^{T,k} = \bar{h}_{xy}\left(\gamma, \mu^{P,k}\right)$. The number of rejected offers at step k from taxis of type x to passengers of type y is thus $\mu_{xy}^{P,k} - \mu_{xy}^{T,k}$; the number of available offers $\mu_{xy}^{A,k}$ in this segment is thus decreased by as much at the end of step k. Formally, the algorithm is described as:

Algorithm 1. *Step 0. Initialize the number of available taxis by*

$$\mu_{xy}^{A,0} = n_x.$$

Step $k \geq 1$. There are three phases:

Proposal phase: Passengers make proposals subject to availability constraint:

$$\mu^{P,k} \in \arg\max_{\mu} \left\{ \sum_{xy} \mu_{xy}\alpha_{xy} - G^*\left(\mu\right) : \mu \leq \mu^{A,k-1} \right\}.$$

Disposal phase: Taxis pick up their best offers among the proposals:

$$\mu^{T,k} \in \arg\max_{\mu} \left\{ \sum_{xy} \mu_{xy}\gamma_{xy} - H^*\left(\mu\right) : \mu \leq \mu^{P,k} \right\}.$$

Update phase: The number of available offers is decreased according to the number of rejected ones

$$\mu^{A,k} = \mu^{A,k-1} - \left(\mu^{P,k} - \mu^{T,k}\right).$$

The algorithm stops when the norm of $\mu^{P,k} - \mu^{T,k}$ is below some tolerance level.

Numerically, the proposal (and disposal) phase requires calculating $\bar{G}$ (defined in eq. (B.4)), and its gradient, which is the constrained demand as if the exogenously quantity cap is $\mu^{A,k-1}$. This amounts to solve an optimal transport problem for which efficient algorithms have been extensively studied.

Appendix E. Proofs

E.1. **Lemmas.**

Lemma 1.

Proof. G is convex as the sum of convex functions, and G^* is convex as the maximum of affine functions. Let us show that G is submodular. One has

$$\frac{\partial G\left(U\right)}{\partial U_{xz}} = \mathbb{E}\left[1\left\{U_{xz}+\varepsilon_z \geq \max_{z'\in\mathcal{Z}_0\backslash\{z\}}\left\{U_{xz'}+\varepsilon_{z'}\right\}\right\}\right].$$

But for $z' \neq z$, the random map

$$U_{xz'} \to 1\left\{U_{xz}+\varepsilon_z \geq \max_{z''\in\mathcal{Z}_0\backslash\{y\}}\left\{U_{xz''}+\varepsilon_{z''}\right\}\right\}$$

is nonincreasing, and thus $U_{xz'} \to \mathbb{E}\left[1\left\{U_{xz}+\varepsilon_z \geq \max_{z''\in\mathcal{Z}_0\backslash\{y\}}\left\{U_{xz''}+\varepsilon_{z''}\right\}\right\}\right] = \partial G\left(U\right)/\partial U_{xz}$ is nonincreasing too. Hence G is submodular. Because G is submodular, the fact that G^* is supermodular now follows from corollary 2.7.3 in Topkis (1998). ∎

Lemma 2.

Proof. By proposition 3, $\bar{g}_{xz} = \partial\bar{G}/\partial\alpha_{xy}$, hence $\partial\bar{g}_{xz}/\partial\bar{\mu}_{x'z'} = \partial^2\bar{G}/\partial\alpha_{xz}\partial\bar{\mu}_{x'z'}$. Similarly, $T_{x'z'} = \partial\bar{G}/\partial\bar{\mu}_{x'z'}$, hence $\partial T_{x'z'}/\partial\alpha_{xz} = \partial^2\bar{G}/\partial\bar{\mu}_{x'z'}\partial\alpha_{xz}$. Identity (B.14) then follows from Schwarz's theorem. ∎

Lemma 3.

Proof. One has

$$\alpha_{xz} - T_{xz}\left(\alpha,\bar{\mu}\right) = \arg\max_{U\leq\alpha}\left\{-G\left(U\right)+\sum_{xy}\left(U_{xy}-\alpha_{xy}\right)\bar{\mu}_{xy}\right\}.$$

The function $\hat{G}$, defined by $\hat{G}\left(U,\alpha\right) = -G\left(U\right)+\sum_{xy}\left(U_{xy}-\alpha_{xy}\right)\bar{\mu}_{xy}$, is supermodular because G is submodular. Further, the set-valued map $\alpha \to \{U : U \leq \alpha\}$ is increasing. By Topkis' theorem again, $\arg\max_U \hat{G}\left(U,\alpha\right)$ is an isotone function of α. ∎

E.2. **Propositions.**

Proposition 1.

Proof. From expression (B.2), it follows that $\bar{G}\left(\alpha,\bar{\mu}\right)$ can be expressed as

$$\bar{G}\left(\alpha,\bar{\mu}\right)=\min_{\tau\geq 0}\max_{\mu\geq 0}\left\{\sum_{x\in\mathcal{X},z\in\mathcal{Z}}\tau_{xz}\bar{\mu}_{xz}+\sum_{x\in\mathcal{X},z\in\mathcal{Z}}\mu_{xz}\left(\alpha_{xz}-\tau_{xz}\right)-G^{*}\left(\mu\right)\right\}$$

and because the maximum can be restricted to the set of μ's such that $0\leq\mu_{xz}\leq\bar{\mu}_{xz}$, which is compact, while the Lagrangian is concave in μ and convex in τ, it follows that

$$\begin{aligned}\bar{G}\left(\alpha,\bar{\mu}\right) &= \max_{\mu\geq 0}\min_{\tau\geq 0}\left\{\sum_{x\in\mathcal{X},z\in\mathcal{Z}}\tau_{xz}\bar{\mu}_{xz}+\sum_{x\in\mathcal{X},z\in\mathcal{Z}}\mu_{xz}\left(\alpha_{xz}-\tau_{xz}\right)-G^{*}\left(\mu\right)\right\}\\ &= \max_{\mu\geq 0}\left\{\sum_{x\in\mathcal{X},z\in\mathcal{Z}}\mu_{xz}\alpha_{xz}-G^{*}\left(\mu\right)\ s.t.\ \mu_{xz}\leq\bar{\mu}_{xz},\ x\in\mathcal{X},\ z\in\mathcal{Z}\right\}.\end{aligned}$$

■

Proposition 2.

Proof. Let $\mu_{z|x}=\mu_{xz}/n_{x}$. From (B.4) , one has

$$\begin{aligned}\bar{G}\left(\alpha,\bar{\mu}\right)=\max_{\mu\geq 0} &\quad \left\{\sum_{x\in\mathcal{X},z\in\mathcal{Z}}\mu_{xz}\alpha_{xz}+\sum_{x\in\mathcal{X}}n_{x}\max_{\substack{Z\sim\mu_{z|x}\\ \varepsilon\sim\mathbf{P}_{x}}}\mathbb{E}\left(\varepsilon_{z}\right)\right\},\\ s.t. &\quad \mu_{xz}\leq\bar{\mu}_{xz},\ x\in\mathcal{X},\ z\in\mathcal{Z}\end{aligned}$$

and thus

$$\bar{G}\left(\alpha,\bar{\mu}\right)=\max_{\mu_{xz}\leq\bar{\mu}_{xz}}\sum_{x\in\mathcal{X}}n_{x}\{\sum_{z\in\mathcal{Z}}\mu_{z|x}\alpha_{xz}+\max_{\substack{Z\sim\mu_{z|x}\\ \varepsilon\sim\mathbf{P}_{x}}}\mathbb{E}\left(\varepsilon_{z}\right)=\max_{\mu_{xz}\leq\bar{\mu}_{xz}}\sum_{x\in\mathcal{X}}n_{x}\max_{\substack{Z\sim\mu_{z|x}\\ \varepsilon\sim\mathbf{P}_{x}}}\mathbb{E}\left(\alpha_{xZ}+\varepsilon_{z}\right)$$

■

Proposition 3.

Proof. It follows from expression (B.4) that the Legendre-Fenchel transform of $\alpha \to \bar{G}(\alpha, \bar{\mu})$ is $G^*(\mu) + A(\mu; \bar{\mu})$, where $A(\mu; \bar{\mu}) = 0$ if $0 \leq \mu_{xz} \leq \bar{\mu}_{xz}$ for all x and z, and $A(\mu; \bar{\mu}) = +\infty$ otherwise. Under Assumption 1, G^* is stricly convex on the set of μ such that $0 < \mu_{xz} < \bar{\mu}_{xz}$ for all x and z. By theorem 11.13 in Rockafellar and Wets (2009), it follows that its Legendre-Fenchel transform $\alpha \to \bar{G}(\alpha, \bar{\mu})$ is continuously differentiable. By the envelope theorem in (B.4), we get that $\bar{g}_{xz}(\alpha, \bar{\mu}) = \partial \bar{G}(\alpha, \bar{\mu}) / \partial \alpha_{xz}$.

It follows from expression (B.2) that the Legendre-Fenchel transform of $\bar{\mu} \to -\bar{G}(\alpha, \bar{\mu})$ is $G(\alpha + \delta) + B(\delta)$, where $B(\delta) = 0$ if $\delta_{xz} \leq 0$ for all x and z, and $B(\delta) = +\infty$ otherwise. Under assumption 1, $\delta \to G(\alpha + \delta)$ is strictly convex, and thus for the same reasons as above, one concludes that $\bar{\mu} \to \bar{G}(\alpha, \bar{\mu})$ is continuously differentiable. By the envelope theorem in (B.2), we get that $T_{xz}(\alpha, \bar{\mu}) = \partial \bar{G}(\alpha, \bar{\mu}) / \partial \bar{\mu}_{xz}$. ∎

Proposition 4.

Proof. Because of the assumption made on $\mathbf{P}_x$ and $\mathbf{Q}_y$, $\mu_{xy} > 0$ for every x and y. Thus $L(\alpha, \gamma, n, m) = 0$ if and only if $\tau^{\alpha}_{xy} = 0$ and $\tau^{\gamma}_{xy} = 0$ for every x and y. Hence, if $L(\alpha, \gamma, n, m) = 0$, then $\mu = \nabla G(\alpha) = \nabla H(\gamma)$, which implies $\alpha + \gamma = \nabla G^*(\mu) + \nabla H^*(\mu) = \nabla \mathcal{E}(\mu)$, which is the first order condition associated to the strictly concave maximization problem (C.2). Conversely, if μ is a solution of (C.2), then it follows that $\alpha + \gamma = \nabla G^*(\mu) + \nabla H^*(\mu)$; but $\alpha - \tau^{\alpha} = \nabla G^*(\mu)$ and $= \nabla H^*(\mu)$, and thus by summation,

$$\alpha + \gamma - \tau^{\alpha} - \tau^{\gamma} = \nabla G^*(\mu) + \nabla H^*(\mu) = \alpha + \gamma.$$

As τ^{α} and τ^{γ} have nonnegative entries, this implies that they are $\tau^{\alpha} = \tau^{\gamma} = 0$, and thus $L(\alpha, \gamma, n, m) = 0$. ∎

E.3. **Theorems.**

Theorem 1.

Proof. Part (i). Clearly, one has $\mu_{xy} \geq 0$ and $\sum_{y \in \mathcal{Y}} \mu_{xy} \leq n_x$ and $\sum_{x \in \mathcal{X}} \mu_{xy} \leq m_y$. By definition of a stable matching in the classical sense, we have

$$\max\left\{u_i^\mu - \alpha_{ij}, v_j^\mu - \gamma_{ij}\right\} \geq 0.$$

Thus when taking the minimum over the set of i such that $x_i = x$ and the set of j such that $y_j = y$, we get

$$\max\left\{u_x - \alpha_{xy}, v_y - \gamma_{xy}\right\} \geq 0,$$

where u_x and v_y are given by (2.6). Now assume $\mu_{xy} > 0$. Then there is a i and a j such that $x_i = x$ and $y_j = y$ and $\mu_{ij} > 0$. Thus $u_i^\mu = \alpha_{ij}$ and $v_j^\mu = \gamma_{ij}$, hence $\max\left\{u_i^\mu - \alpha_{ij}, v_j^\mu - \gamma_{ij}\right\} = 0$. Thus

$$\max\left\{u_x - \alpha_{xy}, v_y - \gamma_{xy}\right\} = \min_{\substack{i : x_i = 0 \\ j : y_j = 0}} \max\left\{u_i^\mu - \alpha_{ij}, v_j^\mu - \gamma_{ij}\right\} = 0.$$

Part (ii). Assume that (μ, u, v) is an aggregate stable matching with money burning, and consider any vector $\left(\mu_{ij}\right) \in \{0,1\}^{\mathcal{I} \times \mathcal{J}}$ such that

$$\mu_{xy} = \sum_{i \in \mathcal{I}, j \in \mathcal{J}} \mu_{ij} 1\left\{x_i = x\right\} 1\left\{y_j = y\right\}.$$

Let us show that there cannot be any blocking pair. Assume that there is a blocking pair ij and call x and y the respective types of i and j. Then we have that

$$\max\left\{u_i^\mu - \alpha_{xy}, v_j^\mu - \gamma_{xy}\right\} < 0.$$

But we have that there is some $j' \in \mathcal{J}_0$ such that $\mu_{ij'} > 0$, thus, denoting the type of j' as y', we have $\mu_{xy'} > 0$, hence $\max\left\{u_x - \alpha_{xy'}, v_{y'} - \gamma_{xy'}\right\} = 0$, thus $u_x \leq \alpha_{xy'} = u_i^\mu$, and similarly there is some $i' \in \mathcal{I}_0$ of type x' such that $v_y \leq \gamma_{x'y} = v_j^\mu$. Hence

$$\max\left\{u_x - \alpha_{xy}, v_y - \gamma_{xy}\right\} \leq \max\left\{u_i^\mu - \alpha_{xy}, v_j^\mu - \gamma_{xy}\right\} < 0,$$

which is a contradiction. Similarly, let us show that there cannot be any blocking single agent. Assume w.l.o.g. that i blocks μ. Then $u_i^\mu < \alpha_{i0}$. But then as above we can show that $u_x \leq u_i^\mu < \alpha_{i0}$, which contradicts $u_x \geq 0$. ∎

Theorem 2. The proof of theorem 2 is based on the fact that algorithm 1 converges. This convergence itself follows from a series of claims. All these claims assume as in theorem 2 that the distributions $\mathbf{P}_x$ and $\mathbf{Q}_y$ have a non-vanishing density.

Claim 1. *Tentatively accepted offers remain in place at the next period: $\mu^{T,k} \leq \mu^{P,k+1}$.*

Proof. By theorem 2, $\mu^{A,k} \leq \mu^{A,k-1}$ implies $\mu^{A,k} - \mu^{P,k+1} \leq \mu^{A,k-1} - \mu^{P,k}$, thus $\mu^{A,k} - \mu^{A,k-1} + \mu^{P,k} \leq \mu^{P,k+1}$. Thus, $\mu^{T,k} \leq \mu^{P,k+1}$. ∎

Claim 2. *As k grows, $\tau^{G,k}$ weakly increases and $\tau^{H,k}$ weakly decreases.*

Proof. One has $\mu_{xy}^{A,k-1} \leq \mu_{xy}^{A,k}$, thus as ∇G^* is isotone, $\nabla G^* \left(\mu^{A,k-1}\right) \leq \nabla G^* \left(\mu^{A,k}\right)$, hence $\alpha_{xy} - \tau_{xy}^{G,k-1} \leq \alpha_{xy} - \tau_{xy}^{G,k}$. To see that $\tau^{H,k} \geq \tau^{H,k-1}$, note that

$$\tau^{H,k} = T^H \left(\gamma, \mu^{T,k}\right)$$
$$\tau^{H,k+1} = T^H \left(\gamma, \mu^{P,k+1}\right)$$

and $\mu^{T,k} \leq \mu^{P,k+1}$ along with the fact that $T^H \left(\gamma, \bar{\mu}\right)$ is antitone in $\bar{\mu}$ (theorem 5) allows to conclude. ∎

Claim 3. *At every step k, $\min\left(\tau_{xy}^{G,k}, \tau_{xy}^{H,k}\right) = 0$ for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.*

Proof. $\tau_{xy}^{H,k} > 0$ implies $\tau_{xy}^{H,l} > 0$ for $l \in \{1, ..., k\}$; hence $\mu_{xy}^{P,l} = \mu_{xy}^{T,l}$, hence $\mu_{xy}^{A,k-1} = \mu_{xy}^{A,0} = n_x$. Assume $\tau_{xy}^{G,k} > 0$. Then it means that the corresponding constraint is saturated, which means $\mu_{xy}^{P,k} = \mu_{xy}^{A,k-1} = n_x$, a contradiction as $\mu_{xy}^{P,k}$ is necessarily less than n_x because by assumption, $\mathbf{P}_x$ has a non-vanishing density. ∎

Claim 4. *As $k \to \infty$, $\lim \nabla G\left(\alpha - \tau^{G,k}\right) = \lim \nabla H\left(\gamma - \tau^{H,k}\right) =: \mu$. As a result, algorithm 1 converges.*

Proof. One has $\mu^{A,k-1} - \mu^{A,k} = \mu^{P,k} - \mu^{T,k} = \nabla G\left(\alpha - \tau^{G,k}\right) - \nabla H\left(\gamma - \tau^{H,k}\right)$, but as $\mu^{A,k}$ is non-increasing and bounded, this quantity tends to zero. Further, $\tau^{G,k}$ and $\tau^{H,k}$ converge monotonically, which shows that $\lim_k \nabla G\left(\alpha - \tau^{G,k}\right) = \lim_k \nabla H\left(\gamma - \tau^{H,k}\right)$. ∎

We are now ready to prove theorem 2.

Proof of theorem 2. Define $\tau_{xy}^{\alpha} = \lim_{k\to\infty} \tau_{xy}^{\alpha,k}$ and $\tau_{xy}^{\gamma} = \lim_{k\to\infty} \tau_{xy}^{H,k}$. Because of claim 3, one has $\min\left(\tau_{xy}^{\alpha}, \tau_{xy}^{\gamma}\right) = 0$; because of claim 4 and the continuity of ∇G and ∇H, one has

$\nabla G\left(\alpha-\tau^{\alpha}\right)=\nabla H\left(\gamma-\tau^{\gamma}\right)$. Letting μ be this common vector, it follows that $\left(\tau_{xy}^{\alpha},\tau_{xy}^{\gamma}\right)$ is an equilibrium matching under non-price rationing. ∎

Theorem 3.

Proof. Let $F\left(\tau\right)=-e\left(\tau\right)$ where e is defined is defined in (3.8). We would like to show that F is an M-function using the terminology of Rheinboldt (1974). F is an M-function if and only if it satisfies both following properties:

(i) F is off-diagonally isotone: for $xy\neq x'y'$, $F_{xy}\left(\tau\right)$ should be non-increasing in $\tau_{x'y'}$, and

(ii) F is a P-function: for any $\tau\neq\tau'$, there exists x and y (which may depend on τ and τ') such that $\left(\tau_{xy}-\tau'_{xy}\right)\left(F_{xy}\left(\tau\right)-F_{xy}\left(\tau'\right)\right)>0$.

Requirement (i) easily follows from the submodularity of G and H, and the fact that $\tau\to\tau^{+}$ and $\tau\to\tau^{-}$ are respectively isotone and antitone.

Let us show that requirement (ii) is also satisfied. By contradiction, assume that there are price vectors τ and τ' such that $\tau\neq\tau'$ and for all x and y,

$$\left(\tau_{xy}-\tau'_{xy}\right)\left(F_{xy}\left(\tau\right)-F_{xy}\left(\tau'\right)\right)\leq 0.$$

By the submodularity of G and H, it follows that $\sum_{xy}F_{xy}\left(\tau\right)$ should be strictly isotone in $\tau_{x'y'}$ for any x' and y'. Hence, $\sum_{xy}F_{xy}\left(\tau\wedge\tau'\right)<\sum_{xy}F_{xy}\left(\tau\vee\tau'\right)$, thus

$$\sum_{xy:\tau_{xy}>\tau'_{xy}}F_{xy}\left(\tau\wedge\tau'\right)+\sum_{xy:\tau_{xy}\leq\tau'_{xy}}F_{xy}\left(\tau\wedge\tau'\right)<\sum_{xy:\tau_{xy}\geq\tau'_{xy}}F_{xy}\left(\tau\vee\tau'\right)+\sum_{xy:\tau_{xy}<\tau'_{xy}}F_{xy}\left(\tau\vee\tau'\right).$$

The following four statements follow from the fact that F is off-diagonal isotone:

If $\tau_{xy}>\tau'_{xy}$ then $\left(\tau\wedge\tau'\right)_{xy}=\tau'_{xy}$ and $F_{xy}\left(\tau'\right)\leq F_{xy}\left(\tau\wedge\tau'\right)$;

If $\tau_{xy}\leq\tau'_{xy}$ then $\left(\tau\wedge\tau'\right)_{xy}=\tau_{xy}$ and $F_{xy}\left(\tau\right)\leq F_{xy}\left(\tau\wedge\tau'\right)$;

If $\tau_{xy}\geq\tau'_{xy}$, then $\left(\tau\vee\tau'\right)_{xy}=\tau_{xy}$ and $F_{xy}\left(\tau\vee\tau'\right)\leq F_{xy}\left(\tau\right)$;

If $\tau_{xy}<\tau'_{xy}$, then $\left(\tau\vee\tau'\right)_{xy}=\tau'_{xy}$ and $F_{xy}\left(\tau\vee\tau'\right)\leq F_{xy}\left(\tau'\right)$.

Hence, (E.3) implies that

$$\sum_{xy:\tau_{xy}>\tau'_{xy}}F_{xy}\left(\tau'\right)+\sum_{xy:\tau_{xy}\leq\tau'_{xy}}F_{xy}\left(\tau\right)<\sum_{xy:\tau_{xy}\geq\tau'_{xy}}F_{xy}\left(\tau\right)+\sum_{xy:\tau_{xy}<\tau'_{xy}}F_{xy}\left(\tau'\right)$$

thus

$$0 < \sum_{xy:\tau_{xy}>\tau'_{xy}} \left(F_{xy}\left(\tau\right) - F_{xy}\left(\tau'\right)\right) + \sum_{xy:\tau_{xy}<\tau'_{xy}} \left(F_{xy}\left(\tau'\right) - F_{xy}\left(\tau\right)\right)$$

but this comes in contradiction with (E.3), which implies that the right hand-side is weakly negative. Hence, F is a P-function, and thus an M-function. According to theorem 9.1 in Rheinboldt (1974), it follows that F is inverse isotone, hence that it is injective. ∎

Theorem 4.

Proof. Let $\sigma_n = 1/n$. The sequences $-\sigma_n \ln \mu_{x0}\left(\sigma_n\right)$ and $-\sigma_n \ln \mu_{0y}\left(\sigma_n\right)$ are valued in $\mathbb{R}_+$; up to a subsequence extraction, one may set $u_x = -\lim_{n\to+\infty} \sigma_n \ln \mu_{x0}\left(\sigma_n\right) \in \mathbb{R}_+ \cup \{+\infty\}$ and $v_y = -\lim_{n\to+\infty} \sigma_n \ln \mu_{0y}\left(\sigma_n\right) \in \mathbb{R}_+ \cup \{+\infty\}$. Up to further sequence extractions, one may define μ^*_{x0}, μ^*_{0y} and μ^*_{xy} as the respective limits of $\mu_{x0}\left(\sigma_n\right)$, $\mu_{0y}\left(\sigma_n\right)$, and $\mu_{xy}\left(\sigma_n\right)$.

Assume $\mu^*_{x0} > 0$. Then $-\sigma_n \ln \mu_{x0}\left(\sigma\right) \approx -\sigma_n \ln \mu^*_{x0} \to 0$ as $n \to +\infty$, thus $u_x = 0$. Similarly, $\mu^*_{0y} > 0$ implies that $v_y = 0$. We have

$$\begin{aligned} \max\left(u_x - \alpha_{xy}, v_y - \gamma_{xy}\right) &= \lim_{n\to+\infty} \max\left(-\sigma_n \ln \mu_{x0}\left(\sigma_n\right) - \alpha_{xy}, -\sigma_n \ln \mu_{0y}\left(\sigma_n\right) - \gamma_{xy}\right) \\ &= \lim_{n\to+\infty} \left\{-\sigma_n \ln\left(\mu_{xy}\left(\sigma_n\right)\right)\right\} \end{aligned} \tag{E.1}$$

hence as $\mu_{xy}\left(\sigma_n\right)$ is bounded above, the limit is either nonnegative or $+\infty$. Thus

$$\max\left(u_x - \alpha_{xy}, v_y - \gamma_{xy}\right) \geq 0. \tag{E.2}$$

Assume $\mu^*_{xy} > 0$. Then $\sigma_n \ln\left(\mu_{xy}\left(\sigma_n\right)\right) \to 0$ as $n \to +\infty$, and therefore inequality (E.2) is saturated. Thus (μ^*, u, v) satisfy conditions (ii) to (vi) of definition 1, but not necessarily condition (i), as the integrality of μ^*_{xy} is not guaranteed. But by the Birkhoff-von Neumann theorem, μ^*_{xy} is a convex combination of some number K of integral vectors μ^k that satisfy conditions (i) to (iii): $\mu^* = \sum_{k=1}^K w_k \mu^k_{xy}$, where $w_k \geq 0$ and $\sum_{k=1}^K w_k = 1$. Let us show that $\left(\mu^1, u, v\right)$ satisfy conditions (i) to (vi) of definition 1. Conditions (i) to (iii) have been checked, and so have the inequalities in conditions (iv) to (vi); the only condition remaining to be satisfied is the equality case in (iv) to (vi); but $\mu^1_{x0} > 0$ implies $\mu^*_{x0} > 0$, and similarly, $\mu^1_{0y} > 0$ implies $\mu^*_{0y} > 0$ and $\mu^1_{xy} > 0$ implies $\mu^*_{xy} > 0$. ∎

Theorem 5.

Proof. By proposition 1, $\bar{G}\left(\alpha,\bar{\mu}\right)=\min_{\tau\geq 0}\left\{G\left(\alpha-\tau\right)+\sum_{xy}\tau_{xy}\bar{\mu}_{xy}\right\}$, hence

$$\tau=\arg\max_{\tau\geq 0}\left\{-G\left(\alpha-\tau\right)+\sum_{xy}\tau_{xy}\left(-\bar{\mu}_{xy}\right)\right\}.$$

The function $\tilde{G}$ defined by $\tilde{G}\left(\tau,\theta\right)=-G\left(\alpha-\tau\right)+\sum_{xz}\tau_{xz}\theta_{xy}$ is supermodular because G is submodular. By Topkis' theorem (theorem 2.8.1 in Topkis 1998), $T\left(\alpha,\bar{\mu}\right)$ which expressed as $\arg\max_{\tau\geq 0}\tilde{G}\left(\tau,-\bar{\mu}\right)$ is an isotone function of $-\bar{\mu}$, hence it is an antitone function of $\bar{\mu}$. ∎

Theorem 6.

Proof. Because of lemma 3, $\partial\left(\alpha_{x'z'}-T_{x'z'}\left(\alpha,\bar{\mu}\right)\right)/\partial\alpha_{xz}\geq 0$ for every x,x',z,z'. It follows from lemma 2 that $\partial\left(\bar{\mu}_{xz}-\bar{g}_{xz}\left(\alpha,\bar{\mu}\right)\right)/\partial\bar{\mu}_{x'z'}=\partial\left(\alpha_{x'z'}-T_{x'z'}\left(\alpha,\bar{\mu}\right)\right)/\partial\alpha_{xz}$, hence $\partial\left(\bar{\mu}_{xz}-\bar{g}_{xz}\left(\alpha,\bar{\mu}\right),\right)/\partial\bar{\mu}_{x'z'}\geq 0$ for every x,x',z,z', and thus $\bar{\mu}-\bar{g}\left(\alpha,\bar{\mu}\right)$ is isotone in $\bar{\mu}$. ∎