A THEORY OF DECENTRALIZED MATCHING MARKETS WITHOUT TRANSFERS, WITH AN APPLICATION TO SURGE PRICING

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

Abstract. Most of the literature on two-sided matching markets without transfers focuses on the case where a central planner (often an algorithm) clears the market, like in the case of school assignments, or medical residents. In contrast, we focus on decentralized matching markets without transfers, where prices are regulated and thus cannot clear the market, as in the case of taxis. In these markets, time waited in line often plays the role of a numéraire. We investigate the properties of equilibrium in these markets (existence, uniqueness, and welfare). We use this analysis to study the problem of surge pricing: given beliefs on random demand and supply, how should a market designer set prices to minimize expected market inefficiency?

Keywords: Two-sided matching; non-transferable utility matching; rationing by waiting; non-price rationing; disequilibrium; surge pricing; discrete choice.

JEL Classification: C78, D58, D61.

Date: April 10th, 2017. 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, Carnergie Mellon University, Tepper School of Business, CalTech, UNC, Columbia University, Celebrating Chris Flinn's 65th Birthday Conference, 2015 California Econometrics Conference.

1. Introduction

The literature on matching markets usually distinguishes between models with transfers, where a numéraire good–often money–clears the market; and models without transfers, where no exchange of numéraire is possible or allowed. Before the emergence of online matching platforms, models with transfers usually applied to *decentralized* markets such as the labor market, the housing market, or the marriage market. Models without transfers were typically used to represent *centralized* markets such as school assignments, organ transplants, and medical residents. In these centralized markets, in the absence of prices, a market designer–which often consists of a computer–clears the market by the means of an algorithmic procedure.

While most markets with transfers are decentralized and most markets without transfers are centralized, there are some important exceptions. The market of personal transportation provides a notable example of a *decentralized matching market without transfers* in the case of taxis (where the unit fare is fixed by the regulator[1]), and of a *centralized market with transfers* in the case of emerging ride-sharing services such as Uber (where a central computer generates surge prices that vary according to supply and demand, and matches passengers to drivers). This brings two questions which are at the core of this paper: (1) one pertaining to equilibrium with non-price rationing: what clears the market in the case of decentralized matching markets without transfers? and (2) another one on surge pricing: how to set the prices in the case of centralized markets when demand and supply is uncertain?

When prices do not have the freedom to adjust to clear the market, other types of rationing mechanisms emerge, which is the focus of our first question above. Non-price rationing mechanisms include stochastic rationing, black market, and rationing by waiting. Queues, traffic congestion and waiting lines are instances of rationing by waiting on which we will concentrate. In these cases, fixed prices induce agents to wastefully compete by waiting

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

in line[2] in order to get access to a good or to a match. In many cities, there are waiting zones where either passengers wait for taxis, or taxis wait for passengers. Time replaces money as a bidding device: agents with the highest willingness to wait will end up getting matched. However, unlike money, time waited in line is not transferable to the other side of the market: picking up a passenger who has waited for a long time does not benefit a taxi driver. Recently, a number of ride service companies have implemented a surge pricing mechanism, which adjusts supply and demand to minimize waiting lines. The study the surge pricing problem is our second focus: given uncertain demand and supply, and given the need to set prices in advance for some window of time, the question is, which price should one set in order to minimize the total inefficiency resulting from time waited in line.

In this paper, we propose a model of decentralized matching with rationing-by-waiting. Agents have differentiated utilities for being matched to an agent on the other side of the market. Our model has *nontransferable utility* in the sense that no transfer can be made from one side of the market to the other. However, it is a *competitive model* in the sense that waiting lines will form in front of overdemanded agents; and the utility of someone matching with an overdemanded agent will be decreased in proportion to the amount of time waited. In this context, we define a concept of matching equilibrium with rationing-by-waiting; such an equilibrium specifies the matching patterns, as well as the waiting times, if any, associated with each agent. We show the existence of an equilibrium which can be computed by a modification of the deferred acceptance algorithm, and we study the issue of uniqueness of an equilibrium in a large market. We investigate the surge pricing problem, which consists of minimizing the expected measure of market inefficiency when demand and supply are not perfectly forecasted. While most of our paper is written with a random utility term, its last section investigates the case fully deterministic utilities, and makes a precise connection with the classical theory of stable matchings by Gale and Shapley.

[2]Wasteful competition may take other forms than waiting lines, such as fighting or overinvesting.

Our paper is related to three topics in the economics literature: (i) non-price rationing, (ii) surge pricing, and (iii) decentralized matching without transfers, which are reviewed here in order.

First, there is an important prior economic literature on *non-price rationing*, a central issue in economics, which arises in many diverse situations such as sticky prices in the macroeconomic theory of "disequilibrium," see e.g. Bénassy (1976), Gourieroux and Laroque (1985), and Drèze (1987); in credit rationing, see Sealy (1979), in housing market with rent control, see Glaeser and Luttmer (2003), and in health economics, see for example, Lindsay and Feigenbaum (1984), Iversen (1993), Martin and Smith (1999), and a recent survey by Iversen and Siciliani (2011). To the best of our knowledge, the phrase "rationing-by-waiting" was coined by Barzel (1974). The mathematics of queuing are surveyed exposed 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). Recently, waiting lines have been studied from a learning context by Margaria (2016). 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, see Sandel (2014).

Second, *surge pricing* (or dynamic pricing) issues have also abundantly been addressed in the economic literature under various names. In electricity markets, the peak-load pricing problem is a classic problem in public economics, see Williamson (1966); more recently, the introduction of "smart meters" opens up the possibility for utility companies to impose peak-time pricing on their household clients, see Joskow and Wolfram (2012). Surge pricing is also found in the complex dynamic pricing system implemented in the airline and hotel industries, see McAfee and te Velde (2006). A literature on of personal transportation services has emerged since Arnott (1996). Levin and Skrzypacz (2016) focus on competition and externalities whithin and between platforms. Fréchette et al. (2016), and Buchholz (2016) provide a general equilibrium analysis of the taxi market. Castillo et al. (2017) study the effect of hypercongestion and provide recommendations for dynamic pricing.

Finally, there is a large literature on centralized matching models without transfers, called matching design problems, which we shall not review here; we focus instead on the narrower literature on *decentralized matching without transfers*. The basic observation is that it is very hard to define aggregate stable matchings when agents are clustered into types of indistinguishable individuals. Indeed, in the absence of transfers, it can be hard to break ties between identical individuals[3], and therefore it may 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 main directions. First, by stochastic rationing, see Gale (1996) and references therein or by the introduction of search frictions, see e.g. Burdett and Coles (1997), Smith (2006) and the references therein. Search fictions provide a way to ration demand and supply stochastically, and provides a rationale for the variation in the equilibrium payoffs of similar individuals. Second, by the introduction of heterogeneity, which can either be observed, as in a recent paper by Azevedo and Leshno (2016), or can be unobserved, and be captured in a random utility model, such as Dagsvik (2000) and Menzel (2015) who have logit heterogeneities. Recently, Che and Koh (2016) have investigated the case of decentralized college admission with uncertain student preferences. See also Echenique and Yariv (2013), and Niederle and Yariv (2009) for other approaches to study decentralized matching markets. Echenique et al. (2013) hace a characterization of rationalizability of matchings without transfers in the spirit of revealed preference.

Our paper opens up a third approach. By relying on the concept of rationing-by-waiting, we are able to deal with the difficult problem of aggregation in nontransferable matching models. Waiting times play a somewhat similar role as transfers, in the sense that they adjust supply and demand, so that at equilibrium, everyone is happy *unconditionally* with their assignment– in contrast to the classical notion of stability with Non-transferable utility (NTU) of Gale and Shapley (1962), where, at equilibrium, every agent achives their best option only within the pool of potential partners who rank him or her above their current match. Similar to the models of Dagsvik (2000) and Menzel (2015) which are based on

[3]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.

the classical notion of stability, our model allows for stochastic utility components. Yet the incorporation of waiting times allows us to explore a richer structure of distributions of stochastic utilities beyond the logit case considered in these papers. Like Azevedo and Leshno (2016), we can think of our notion of equilibrium as the solution of a tâtonnement process in a demand and supply framework; however, in contrast to that paper, our framework accommodates a finite number of agents, and does not necessitate to consider a continuous limit. Our paper can be seen as the NTU counterpart of the model with transferable utility and stochastic utility of Galichon and Salanié (2016), extending the Choo and Siow (2006) model beyond the logit case. The present model can be seen as a the limiting case of models with imperfectly transferable utility and stochastic utility of Galichon et al. (2016); however, a detailed study of the limiting case was left aside in that paper for further investigations.

The present paper's contribution is two-fold. First, it offers a framework for decentralized NTU equilibria. This leads us to propose a definition of NTU stability which is related to the classical one, but not identical. Unlike the classical notion of NTU stability, our proposal is suitable for the description of decentralized equilibria because it is based on an explicit competitive rationing mechanism. It also permits a very natural definition of aggregate stable matchings, where an equilibrium exists and is unique. Our framework is easily taken to data, as it can be interfaced with a random utility model, becoming a NTU equivalent to the model of Choo and Siow (2006) in the logit case and beyond that case. Because of its tractable computational properties, our model is well suited for welfare analysis, a natural measure of welfare loss being the total amount of time waited in line on both sides of the market. As a by-product, this welfare analysis allows us to build a surge pricing algorithm, which minimizes welfare loss when market parameters, such as supply and demand, are uncertain. A second, distinct, contribution of this paper, is to develop a theory of random utility under rationing constraints. In many settings where goods are available in fixed supply, but where prices are rigid, congestion parameters, such as waiting time, serves as a market clearing numéraire. With more and more abundant

data on everything, there are more and more empirical settings where such waiting times become available to the econometrician, making such tools empirically relevant.

The rest of the paper is organized as follows. Section 2 provides a bird's-eye view of our approach. Section 3 provides a brief review of classical discrete choice theory (without rationing). Section 4 incorporates rationing-by-waiting into this classical theory. Section 5 introduces the matching setting, and defines a notion of equilibrium and solve for it. Section 6 makes the link with the classical theory of stable matchings without transfers à la Gale and Shapley.

2. The problem in a nutshell

A desirable feature of an equilibrium concept is that two identical agents should get identical payoffs at equilibrium. A notable violation of this requirement, however, is exemplified by the notion of stable matchings without transfers, where it is easy to come up with an example where there is no single stable outcome in which two identical agents will obtain the same payoff at equilibrium. Here is a simple example:

Example 1. *Assume that there are 2 identical passengers and 1 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 a model with transfers, if the taxi picks up a passenger, then the total surplus is 2, and because an identical passenger is unmatched, the matched passenger will get utility zero (otherwise the unmatched passenger could outbid her), and thus the taxi appropriates the total payoff amount of 2.*

In a classical model without transfers, there are two stable matchings in each of which the matched passenger gets one, while the unmatched gets zero. There is no stable matching in which both passengers get the same payoff, thus demand has to be rationed stochastically.

The aim of this paper is to provide a natural notion of equilibrium without transfers where any two identical passengers will end up with the same equilibrium payoff. In a model of matching with rationing-by-waiting, both passengers will compete by waiting in line, or by fighting, until their payoff reaches the reservation value of zero. Thus there are still two

stable matchings, but in both of these, both passengers get zero payoff, while the taxi gets a payoff amount 1 (instead of 2 in the case of a transferable utility model).

Here is a way to represent this rationing mechanism in a supply and demand framework. Assume prices p are free to move around. Let $D(p)$ be the demand, $S(p)$ be the supply. Then the prices adjust to balance demand and supply:

$$Z(p) := D(p) - S(p) = 0, \tag{2.1}$$

which is a very classical Walrasian equilibrium equation.

Now assume that prices are not free to adjust, and are set to a fixed value $\overline{p}$, and hence demand and supply will not clear. A queue (either on the supply or on the demand side) will form. Let $\tau^D \geq 0$ (resp. $\tau^S \geq 0$) be the monetary equivalent of the time waited on the demand (resp. supply) side. With the waiting times, demand is $D\left(\overline{p} + \tau^D\right)$ and supply is $S\left(\overline{p} - \tau^S\right)$. Letting $\tau = \tau^D - \tau^S$ so that $\tau^D = \tau^+$ and $\tau^S = \tau^-$, τ is determined at equilibrium by

$$Z(\tau; \overline{p}) := D\left(\overline{p} + \tau^+\right) - S\left(\overline{p} - \tau^-\right) = 0, \tag{2.2}$$

where again, $\overline{p}$ is fixed, but where it is now for τ to adjust in order to balance demand and supply.

Equilibrium equations (2.1) and (2.2) have both similarities and differences. The solution techniques (tâtonnement) are similar. However, a main difference is that (unlike monetary transfers) the time waited by one side of the market is costly to that side, but is not beneficial to the other side of the market. A measure of market inefficiency will then naturally depend on τ^+, the time waited on the supply side, and on τ^-, the time waited on the demand side. Therefore, the total market inefficiency can be captured by $l^D\left(\tilde{\tau}^+\right) + l^S\left(\tilde{\tau}^-\right)$, where the loss functions l^S and l^D are increasing and pass through zero.

We are now ready to describe the surge pricing problem in our stylized framework. This is the problem of setting the price $\overline{p}$ in order to minimize the measure of market inefficiency. If demand and supply were perfectly known, then choosing $\overline{p}$ solution of (2.1) would yield zero inefficiency. However, in practical situations, the central planner (e.g. Uber's algorithm) only knows an estimate of demand and supply. As a result, excess demand is random;

denote it $\tilde{Z}(\tau;\overline{p})$. The problem is therefore to minimize the *expected market inefficiency*, that is

$$\min_{\overline{p}} \mathbb{E}\left[l^D\left(\tilde{\tau}^+\right) + l^S\left(\tilde{\tau}^-\right)\right]$$
$$s.t.\ \tilde{Z}(\tilde{\tau};\overline{p}) = 0\ a.s.$$

which is a mathematical programming problem under equilibrium constraints (MPEC). In the rest of the paper, we shall explicitly model all these ingredients and study the properties of the resulting objects.

3. Reminders on discrete choice without rationing

3.1. **Demand.** We start by recalling results on discrete choice without rationing. We will adopt the standard random utility framework (see McFadden, 1976, and Fudenberg et al., 2015) 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. This type may encompass some of the geographic, demographic, and socioeconomic characteristics of the agent. Utility U_{xz} is associated to option $z \in \mathcal{Z}$, while the utility associated to the outside option is normalized to zero. The consumer has an additive random utility term ε_z associated to alternative $z \in \mathcal{Z}_0$. We assume that the distribution of the random utility components (ε_z) drawn by an agent of type x is $\mathbf{P}_x$. Throughout most of the paper, we shall make use of the following assumption:

Assumption 1. *For all $x \in \mathcal{X}$, $\mathbf{P}_x$ has a nowhere vanishing density.*

The demand for of alternative $z \in \mathcal{Z}$ from agents of type x, denoted $g_{xz}(U)$, is the number of consumers of type x demanding alternative z, defined by

$$g_{xz}(U) := n_x \mathbf{P}_x\left(U_{xz} + \varepsilon_z \geq \max_{z' \in \mathcal{Z}_0}\{U_{xz'} + \varepsilon_{z'}, \varepsilon_0\}\right)$$

(note that there is zero probability that a consumer will be indifferent between any pair of alternatives). Note that it is assumed that there is an infinite number of agents per type, thus $\mathbf{P}_x(U_{xz} + \varepsilon_z \geq ...)$ describes both the probability and the frequency of the choice of z by an agent of type x. Therefore the demand map g_{xz} is deterministic, and does not depend on a particular realization of the random utility vector ε. It would be random if there were

only a finite number of agents per type, because then the realization of the vector of utility shocks would introduce some randomness.

The map g can be conveniently expressed using the celebrated Daly-Zachary-Williams theorem (see Rust, 1994 and references therein). Introducing the expected indirect utility of the consumer, given by

$$G(U) := \sum_{x \in \mathcal{X}} n_x \mathbb{E}_{\mathbf{P}_x} \left[\max_{z \in \mathcal{Z}_0} \{U_{xz} + \varepsilon_z, \varepsilon_0\} \right],$$

then, by the Daly-Zachary-Williams theorem, $g_{xz}(U)$ obtains as the derivative of $G(U)$ with respect to the systematic utility U_{xz} associated by consumers of type x to alternative z, that is

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

3.2. **Entropy and welfare.** The entropy of choice associated to the consumer's problem, introduced by Galichon and Salanié (2015), hereafter GS, 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 thus given by

$$G^*(\mu) = \sum_{x \in \mathcal{X}} n_x \max_{(U_{xz}) \in \mathbb{R}^{\mathcal{X} \times \mathcal{Z}}} \left\{ \sum_{x \in \mathcal{X}, z \in \mathcal{Z}} \mu_{xz} U_{xz} - G(U) \right\}. \tag{3.1}$$

Following GS, $-G^*(s)$ is interpreted as $\sum_{x \in \mathcal{X}} n_x \mathbb{E}_{\mathbf{P}_x}[\varepsilon_Z]$, where Z is a random variable valued in $\mathcal{Z}_0$ which is the choice of the consumer with a random utility component ε. Thus, we have the breakdown of the expected indirect utility $G(U)$ as

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

Remark 3.1. As shown by GS, proposition 2, we have that

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

hence G^* is the solution to an optimal transport problem.

In some cases, like in the logit case, a closed-form expression exists for G and G^*.

Example 2. *When the random utility components* $(\varepsilon_y)_y$ *are distributed with i.i.d. Gumbel distribution, this model boils down to a standard logit model. In which case, it is well known (see e.g. McFadden 1976) that*

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

and, as shown in GS, G^* *can also be obtained in closed form by*

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

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

3.3. **Comparative statics.**

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

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

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

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

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

is nonincreasing, and thus $U_{xz'} \to \mathbb{E}\left[1\left\{U_{xz} + \varepsilon_z \geq \max_{z'' \in \mathcal{Z}_0 \setminus \{y\}} \{U_{xz''} + \varepsilon_{z''}\}\right\}\right] = \partial G(U)/\partial U_{xz}$ 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). ∎

The economic interpretation of the submodularity of G is by the *gross substitute* property: the demand for z weakly decreases when the systematic utility associated with alternative z' decreases. Of course, $\partial g_{xz}(U)/\partial U_{xz} \geq 0$.

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; this conclusion would follow from the results of Rheinboldt (1974), theorem 9.6, proven under more general assumptions.

4. Discrete choice under rationing-by-waiting

We now extend the analysis to the case when some of the alternatives are subject to a fixed supply constraint. In this case, and in the absence of market-clearing prices, a rationing mechanism needs to take place. The best way to think about our rationing mechanism is using the image of waiting lines. Time waited in line is costly, and serves as a (inefficient) numéraire to clear the market. Let us describe what becomes of the random utility model subject to this rationing constraint.

4.1. **Welfare.** Assume that a maximum number $\bar{\mu}_{xz}$ of consumers of type x can obtain alternative $z \in \mathcal{Z}$. In this case, consumers will compete to obtain the alternative z by waiting in line for it. 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 to alternative z is quasilinear in time waited, and writes $U_{xz} = \alpha_{xz} - \tau_{xz}$. Let μ_{xz} be the number of consumers of type x choosing alternative z. Then τ_{xz} interprets as 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

$$\begin{cases} \mu_{xz} = \partial G\left(\alpha_{xz} - \tau_{xz}\right)/\partial U_{xz} \\ \mu_{xz} \leq \bar{\mu}_{xz}, \tau_{xz} \geq 0 \\ \tau_{xz} > 0 \Longrightarrow \mu_{xz} = \bar{\mu}_{xz}. \end{cases} \tag{4.1}$$

As it turns out, these conditions are the first order conditions associated with a constrained optimization program that generalizes (3.1). 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 $\pi\left(z|x,\varepsilon\right)$ of assigning an invidual of type (x,ε) to an alternative z. Letting $d\pi\left(\varepsilon, z|x\right) = \pi\left(z|x,\varepsilon\right) d\mathbf{P}_x$ be the induced joint distribution on (ε, z) conditional on the type x, the social planner's problem is to pick $\pi\left(\varepsilon, z|x\right)$ so to maximize the overall utility $\sum_{x\in\mathcal{X}} n_x \int \left(\alpha_{xz} + \varepsilon_z\right) d\pi\left(\varepsilon, z|x\right)$ subject to the constraint that under $\pi\left(\varepsilon, z|x\right)$, the distribution of ε is $\mathbf{P}_x$, and the probability number of z is less or equal than $\bar{\mu}_{xz}$. Letting $\overline{\mathcal{M}}_x\left(\mathbf{P}, \bar{\mu}\right)$

be the set of such distributions, formally defined as

$$\overline{\mathcal{M}}_x(\mathbf{P},\bar{\mu}) = \left\{ \pi(\varepsilon,z|x) : \sum_z \pi(\varepsilon,z|x) = \mathbf{P}_x(d\varepsilon)\, n_x \text{ and } \int \pi(\varepsilon,z|x)\, d\varepsilon \leq \bar{\mu}_{xz} \right\}, \quad (4.2)$$

we are in a position to define the *capacity-constrained welfare function* as

$$\bar{G}(\alpha,\bar{\mu}) = \sum_{x\in\mathcal{X}} n_x \max_{\pi(.,.|x)\in\overline{\mathcal{M}}_x(\mathbf{P},\bar{\mu})} \int (\alpha_{xz}+\varepsilon_z)\, d\pi(\varepsilon,z|x). \quad (4.3)$$

Remark 4.1. The computation of $\bar{G}$ amounts to solving $|\mathcal{X}|$ subproblems that are themselves continuous matching problems, that is, optimal transport problem. This fact is very useful in the numerical applications, as it allows for efficient computation of $\bar{G}$ and its gradient.

The following result provides a simple expression of $\bar{G}$ which involves $G^*(\mu)$ defined in (3.1) above.

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

$$\begin{aligned} \bar{G}(\alpha,\bar{\mu}) = \max_{\mu\geq 0} \quad & \left\{ \sum_{x\in\mathcal{X},z\in\mathcal{Z}} \mu_{xz}\alpha_{xz} - G^*(\mu) \right\}. \\ s.t. \quad & \mu_{xz} \leq \bar{\mu}_{xz},\ x\in\mathcal{X},\ z\in\mathcal{Z} \end{aligned} \quad (4.4)$$

Proof. Let $\mu_{z|x} = \mu_{xz}/n_x$. From (4.3), one has

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

and thus

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

and expression (4.4) follows from remark 3.1. ∎

By duality, $\bar{G}$ has a second expression as a function of G, whose interpretation is also interesting.

Proposition 2. *The value of the maximum capacity-constrained social welfare is given by*

$$\bar{G}(\alpha,\bar{\mu}) = \min_{\tau\geq 0}\left\{ G(\alpha-\tau) + \sum_{x\in\mathcal{X},z\in\mathcal{Z}} \bar{\mu}_{xz}\tau_{xz} \right\}. \tag{4.5}$$

Proof. From expression (4.4), it follows that $\bar{G}(\alpha,\bar{\mu})$ can be expressed as

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

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

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

QED. ∎

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

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

where $\bar{G}(\alpha,\bar{\mu})$ interprets 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, we need a price vector τ to clear demand and supply, and the second best welfare actually achieved by consumers is only $G(\alpha-\tau)$. As a result, $\sum_{x\in\mathcal{X},z\in\mathcal{Z}} \bar{\mu}_{xz}\tau_{xz} = \sum_{x\in\mathcal{X},z\in\mathcal{Z}} \mu_{xz}\tau_{xz}$ is the total efficiency loss, which is the total time wasted in line in our interpretation. This quantity is a natural measure of departure from efficiency.

Expressions (4.4) and (4.5) immediately imply the following consequence:

Proposition 3. $\bar{G}(\alpha,\bar{\mu})$ *is convex in* α *and concave in* $\bar{\mu}$.

4.2. **Demand.** In this paragraph, we show that, using function $\bar{G}$ only, we can deduce both the constrained demand $\left(\bar{g}_{xz}\left(\alpha,\bar{\mu}\right)\right)_{xz}$, which is the vector number of consumers of each type choosing an option of each type, and the vector of waiting times $\left(T_{xz}\left(\alpha,\bar{\mu}\right)\right)_{xz}$ waited in each segment of the market. $\bar{g}$ and T are fomally defined as

$$\begin{aligned} \bar{g}\left(\alpha,\bar{\mu}\right) = \text{argmax}_{\mu\geq 0} \quad & \left\{ \sum_{x\in\mathcal{X},z\in\mathcal{Z}} \mu_{xz}\alpha_{xz} - G^{*}\left(\mu\right) \right\}, \\ s.t. \quad & \mu_{xz} \leq \bar{\mu}_{xz},\ x\in\mathcal{X},\ z\in\mathcal{Z} \end{aligned} \tag{4.6}$$

and

$$T\left(\alpha,\bar{\mu}\right) = \text{argmin}_{\tau\geq 0} \left\{ G\left(\alpha-\tau\right) + \sum_{x\in\mathcal{X},z\in\mathcal{Z}} \bar{\mu}_{xz}\tau_{xz} \right\}. \tag{4.7}$$

The fact that these maximizers exist and are unique is a consequence of the following result:

Proposition 4. *Under assumption 1, $\bar{G}$ is continuously differentiable and one has*

$$\bar{g}_{xz}\left(\alpha,\bar{\mu}\right) = \frac{\partial\bar{G}\left(\alpha,\bar{\mu}\right)}{\partial\alpha_{xz}} \text{ and } T_{xz}\left(\alpha,\bar{\mu}\right) = \frac{\partial\bar{G}\left(\alpha,\bar{\mu}\right)}{\partial\bar{\mu}_{xz}}. \tag{4.8}$$

Proof. It follows from expression (4.4) that the Legendre-Fenchel transform of $\alpha\to\bar{G}\left(\alpha,\bar{\mu}\right)$ is $G^{*}\left(\mu\right)+A\left(\mu;\bar{\mu}\right)$, where $A\left(\mu;\bar{\mu}\right)=0$ if $0\leq\mu_{xz}\leq\bar{\mu}_{xz}$ for all x and z, and $A\left(\mu;\bar{\mu}\right)=+\infty$ 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 $\alpha\to\bar{G}\left(\alpha,\bar{\mu}\right)$ is continuously differentiable. By the envelope theorem in (4.4), we get that $\bar{g}_{xz}\left(\alpha,\bar{\mu}\right)=\partial\bar{G}\left(\alpha,\bar{\mu}\right)/\partial\alpha_{xz}$.

It follows from expression (4.5) that the Legendre-Fenchel transform of $\bar{\mu}\to-\bar{G}\left(\alpha,\bar{\mu}\right)$ is $G\left(\alpha+\delta\right)+B\left(\delta\right)$, where $B\left(\delta\right)=0$ if $\delta_{xz}\leq 0$ for all x and z, and $B\left(\delta\right)=+\infty$ otherwise. Under assumption 1, $\delta\to G\left(\alpha+\delta\right)$ is strictly convex, and thus for the same reasons as above, one concludes that $\bar{\mu}\to\bar{G}\left(\alpha,\bar{\mu}\right)$ is continuously differentiable. By the envelope theorem in (4.5), we get that $T_{xz}\left(\alpha,\bar{\mu}\right)=\partial\bar{G}\left(\alpha,\bar{\mu}\right)/\partial\bar{\mu}_{xz}$. ∎

Combining (4.8) with the first order conditions in (4.7), 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 with systematic utility is α under a capacity constraint is $\bar{\mu}$.

Example 2 (continued). *In the logit case, system (4.1) boils down to*

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

hence, it follows from the first equation that $\mu_{xz} = \mu_{x0} \exp\left(\alpha_{xz} - \tau_{xz}\right)$*, thus*

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

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

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

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

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

4.3. **Comparative statics.** We investigate what happens to the waiting times and to the demand when the availability constraint is tightened (i.e. 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 1. *Under assumption 1, the shadow price $T(\alpha, \bar{\mu})$ is an antitone function of the vector of number of available offers $\bar{\mu}$.*

Proof. By proposition 2, $\bar{G}(\alpha, \bar{\mu}) = \min_{\tau \geq 0} \left\{ G(\alpha - \tau) + \sum_{xy} \tau_{xy} \bar{\mu}_{xy} \right\}$, hence

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

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

Theorem 1 can be interpreted as a generalization of 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 (ie when the entries of $\bar{\mu}$ weakly decrease), the number of nondemanded options also weakly decreases in each market segment.

Theorem 2. *Under assumption 1, the number of nondemanded options $\bar{\mu} - \bar{g}(\alpha, \bar{\mu})$, is an isotone function of the capacity vector $\bar{\mu}$.*

The intuition behind theorem 2 is that if the capacity on segment xz is increased, the choices that were dominated are still dominated. This also is 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*

$$\frac{\partial \bar{g}_{xz}}{\partial \bar{\mu}_{x'z'}} = \frac{\partial T_{x'z'}}{\partial \alpha_{xz}}. \tag{4.9}$$

Proof. By proposition 4, $\bar{g}_{xz} = \partial \bar{G}/\partial \alpha_{xy}$, 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 (4.9) then follows from Schwarz's theorem. ∎

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

Example 2 (continued). *In the logit case,* $\partial \bar{g}_{xz}(\alpha, \bar{\mu}) / \partial \bar{\mu}_{xz} = 1\{\bar{\mu}_{xz} \leq \mu_{x0}^* e^{\alpha_{xz}}\}$ *and* $\partial T(\alpha, \bar{\mu}) / \partial \alpha_{xz} = 1\{\alpha_{xy} + \log(\mu_{x0}^*/\bar{\mu}_{xz}) \geq 0\}$*, which obviouly coincide.*

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

Proof. One has

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

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

We are then able to prove theorem 2.

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

5. Matching equilibrium under rationing-by-waiting

5.1. **Equilibrium: existence.** We are ready to define equilibrium in a two-sided matching situation. To make exposition more concrete, we shall adopt the language of passengers and taxi drivers: the passengers characteristics $x \in \mathcal{X}$ may include the location of pickup as well as the size of the party, a possible VIP status, and possibly the destination. The taxi characteristics $y \in \mathcal{Y}$ may include the size of the car, the fleet it belongs to, and the set of amenities offered to passengers, such as A/C, credit card terminal, video screen, etc.

We shall assume there are n_x passengers of type x; a passenger of type x has systematic utility α_{xy} associated with traveling in a car of type y, and a random utility components $\varepsilon \sim \mathbf{P}_x$, whose distribution $\mathbf{P}_x$ may depend on x. Symmetric quantities are defined on the side of drivers. There are m_y drivers of type y. A driver of type y has systematic utility γ_{xy} associated with picking up a passenger of type x, and a utility component $\eta \sim \mathbf{Q}_y$ whose distribution $\mathbf{Q}_y$ may depend on y.

In our setting, everything is known to the market participants, including the the determinstic parts of the utility (the vectors α and γ), the stochastic parts (the realizations of random utility terms ε and η), and the supply of agents of each type (the vectors n and m). We define as before G and G^*, as well and $\bar{G}$, the availability-constrained social welfare function, whose expression is recalled to be:

$$\begin{aligned} \bar{G}(\alpha, \bar{\mu}) = \max_{\mu \geq 0} \quad & \left\{ \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \mu_{xy} \alpha_{xy} - G^*(\mu) \right\} \\ s.t. \quad & \mu_{xy} \leq \bar{\mu}_{xy}, \ x \in \mathcal{X}, \ y \in \mathcal{Y} \end{aligned} \tag{5.1}$$

and under the assumption that $\mathbf{P}_x$ has a nowhere vanishing density, the availability-constrained demand $\bar{g}$ and the demand-side congestion time T^G are deduced from $\bar{G}$ by

$$\bar{g}_{xy}(\alpha, \bar{\mu}) = \frac{\partial \bar{G}}{\partial \alpha_{xy}}(\alpha, \bar{\mu}) \text{ and } T^G_{xy}(\alpha, \bar{\mu}) = \frac{\partial \bar{G}}{\partial \bar{\mu}_{xy}}(\alpha, \bar{\mu}). \tag{5.2}$$

On the drivers' side, the availability-constrained social welfare function $\bar{H}(\gamma, \bar{\mu})$ is defined as

$$\bar{H}(\gamma, \bar{\mu}) = \max_{\mu \geq 0} \quad \left\{ \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \mu_{xy} \gamma_{xy} - H^*(\mu) \right\} \tag{5.3}$$
$$s.t. \qquad \mu_{xy} \leq \bar{\mu}_{xy}, \ x \in \mathcal{X}, \ y \in \mathcal{Y}$$

from which, under the assumption that $\mathbf{Q}_y$ has a nowhere vanishing density, we can deduce the availability-constrained supply $\bar{h}$ and the supply-side congestion time T^H as

$$\bar{h}_{xy}(\gamma, \bar{\mu}) = \frac{\partial \bar{H}}{\partial \gamma_{xy}}(\gamma, \bar{\mu}) \text{ and } T^H_{xy}(\gamma, \bar{\mu}) = \frac{\partial \bar{H}}{\partial \bar{\mu}_{xy}}(\gamma, \bar{\mu}). \tag{5.4}$$

Let us now define our concept of equilibrium. At equilibrium, demand should equal supply, and there cannot be 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. The first constraint is expressed as

$$\mu_{xy} = \bar{g}(\alpha, \mu) = \bar{h}_{xy}(\gamma, \mu) \ \forall x \in \mathcal{X}, y \in \mathcal{Y},$$

while the second one is expressed as

$$\min\left(T^G_{xy}(\alpha, \mu), T^H_{xy}(\gamma, \mu)\right) = 0.$$

Define $\tau^\alpha_{xy} = T^G_{xy}(\alpha, \mu)$ and $\tau^\gamma_{xy} = T^H_{xy}(\gamma, \mu)$ as the vector of time effectively waited for on the demand and supply sides, respectively. The, by the virtues of proposition 4, one gets

$$\bar{g}(\alpha, \mu) = \frac{\partial G}{\partial \alpha_{xy}}(\alpha - \tau^\alpha) \text{ and } \bar{h}_{xy}(\gamma, \mu) = \frac{\partial H}{\partial \gamma_{xy}}(\gamma - \tau^\gamma),$$

thus the equilibrium can be equivalently formulated as a system of equations on the waiting times τ^α and τ^γ, that is

$$\begin{cases} \frac{\partial G}{\partial \alpha_{xy}}(\alpha - \tau^\alpha) = \frac{\partial H}{\partial \gamma_{xy}}(\gamma - \tau^\gamma), \ \forall x \in \mathcal{X}, y \in \mathcal{Y} \\ \min\left(\tau^\alpha_{xy}, \tau^\gamma_{xy}\right) = 0, \ \forall x \in \mathcal{X}, y \in \mathcal{Y}. \end{cases}$$

Therefore, we are led to the following definition:

Definition 1. *An equilibrium outcome under rationing-by-waiting is a vector* $(\mu, \tau^\alpha, \tau^\gamma)$*, where* μ_{xy} *is the number matches of type* xy*,* τ^α_{xy} *is the time that passengers of type* x *have to wait for a taxi of type* y*, while* τ^γ_{xy} *is the time taxis of type* y *have to wait for passengers of type* x*, that verify simultaneously:*

- market clearing: the matches μ *are induced by waiting times* τ^α *and* τ^γ*, that is*

$$\mu_{xy} = \frac{\partial G}{\partial \alpha_{xy}}\left(\alpha - \tau^\alpha\right) = \frac{\partial H}{\partial \gamma_{xy}}\left(\gamma - \tau^\gamma\right),\ \forall x \in \mathcal{X}, y \in \mathcal{Y},\ \textit{and} \tag{5.5}$$

- stability: there is no market xy *where there is a positive waiting time both on the passenger and taxi sides, that is*

$$\min\left(\tau^\alpha_{xy}, \tau^\gamma_{xy}\right) = 0,\ \forall x \in \mathcal{X}, y \in \mathcal{Y}. \tag{5.6}$$

Remark 5.1. Note that it follows from (5.5) that $\mu_{xy} \geq 0$ satisfies

$$\begin{cases} \sum_{y \in \mathcal{Y}} \mu_{xy} \leq n_x\ \forall x \in \mathcal{X} \\ \sum_{x \in \mathcal{X}} \mu_{xy} \leq m_y\ \forall y \in \mathcal{Y} \end{cases} \tag{5.7}$$

where the first set of inequalities is a consequence of $\mu = \nabla G\left(\alpha - \tau^\alpha\right)$, and the second set of inequalities is a consequence of $\mu = \nabla H\left(\gamma - \tau^\gamma\right)$. This is the reason why conditions (5.7) do not appear explicitly in definition 1, as they are implicitly imposed.

We are now going to establish the following theorem:

Theorem 3. *Assume that the distributions* $\mathbf{P}_x$ *and* $\mathbf{Q}_y$ *have a nonvanishing density. Then an equilibrium outcome under rationing-by-waiting exists.*

The proof is constructive, and is based on a modification of the deferred acceptance algorithm of Gale and Shapley (1962). The general principle of this algorithm is well-known; the details need to be adapted to the current setting with discrete choice and rationing-by-waiting. At the initial step of the algorithm, the proposing side of the market, say taxis, makes offers to the other side of the market, say passengers, without any constraint. If some passengers receive a volume of offers that exceeds the capacity they can take, they shall reject their least preferred ones, and tentatively accept the other ones. The volume of rejected offers then subtracts from the total number of offers that are available to taxis;

another rounds begins, where taxis now propose subject to availability constraints. The algorithm loops until all the offers are accepted.

We formalized this idea by introducing the following notation. 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:

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

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

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

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.

The proof of theorem 3 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 3 that the distributions $\mathbf{P}_x$ and $\mathbf{Q}_y$ have a nonvanishing 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 $T^H(\gamma, \bar{\mu})$ is antitone in $\bar{\mu}$ (theorem 1) 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 nonvanishing 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 nonincreasing 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 3.

Proof of theorem 3. 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 outcome under non-price rationing. ∎

5.2. **Equilibrium: uniqueness.** An equilibrium outcome under non-price rationing has been defined as the record of a matching numbers $\mu_{xy} \geq 0$, demand waiting times $\tau_{xy}^{\alpha} \geq 0$, and supply waiting times $\tau_{xy}^{\gamma} \geq 0$ such that

$$\forall x \in \mathcal{X}, y \in \mathcal{Y} \begin{cases} \frac{\partial G}{\partial \alpha_{xy}}\left(\alpha - \tau^{\alpha}\right) = \frac{\partial H}{\partial \gamma_{xy}}\left(\gamma - \tau^{\gamma}\right) \\ \min\left(\tau_{xy}^{\alpha}, \tau_{xy}^{\gamma}\right) = 0 \end{cases} . \tag{5.8}$$

Introducing $\tau_{xy} = \tau_{xy}^{\alpha} - \tau_{xy}^{\gamma}$, so that τ_{xy}^{α} and τ_{xy}^{γ} are respectively the positive part and the negative part of τ_{xy}

$$\tau_{xy}^{\alpha} = \tau_{xy}^{+}, \text{ and } \tau_{xy}^{\gamma} = \tau_{xy}^{-}, \tag{5.9}$$

we can characterize the equilibrium (5.8) as the solution of a system of nonlinear equations

$$E\left(\tau\right) = 0, \tag{5.10}$$

where E the market excess demand function defined by

$$E_{xy}\left(\tau\right) := \frac{\partial G}{\partial \alpha_{xy}}\left(\alpha - \tau^{+}\right) - \frac{\partial H}{\partial \gamma_{xy}}\left(\gamma - \tau^{-}\right). \tag{5.11}$$

The following theorem implies that the solution τ of this equation is unique.

Theorem 4. *Assume that the distributions $\mathbf{P}_x$ and $\mathbf{Q}_y$ have a nonvanishing density. Then the equilibrium outcome under non-price rationing is unique.*

Before we state the proof, let us note that this result is driven by the fact that the distributions of the random utility components are continuous. By contrast we shall study in section 6 the case with fully determinsitic utilities, where there may be multiple equilibria.

Proof. Let $F(\tau) = -E(\tau)$ where E is defined is defined in (5.11). 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 $xy \neq x'y'$, $F_{xy}(\tau)$ should be nonincreasing 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 $\left(\tau_{xy} - \tau'_{xy}\right)\left(F_{xy}(\tau) - F_{xy}(\tau')\right) > 0$.

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,

$$\left(\tau_{xy} - \tau'_{xy}\right)\left(F_{xy}(\tau) - F_{xy}\left(\tau'\right)\right) \leq 0. \tag{5.12}$$

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

$$\sum_{xy:\tau_{xy}>\tau'_{xy}} F_{xy}\left(\tau \wedge \tau'\right) + \sum_{xy:\tau_{xy}\leq\tau'_{xy}} F_{xy}\left(\tau \wedge \tau'\right) < \sum_{xy:\tau_{xy}\geq\tau'_{xy}} F_{xy}\left(\tau \vee \tau'\right) + \sum_{xy:\tau_{xy}<\tau'_{xy}} F_{xy}\left(\tau \vee \tau'\right). \tag{5.13}$$

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

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

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

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

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

Hence, (5.13) implies that

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

thus

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

but this comes in contradiction with (5.12), 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. ∎

5.3. **Welfare.** Next, we would like to evaluate the *congestion inefficiency*, which is due to time waited in line. If a passenger x waits for a taxi of type y an amount of time $\tau_{xy}^{\alpha} \geq 0$, let $l_{xy}^{G}\left(\tau_{xy}^{\alpha}\right)$ be the corresponding social loss; similarly, let us denote $l_{xy}^{H}\left(\tau_{xy}^{\gamma}\right)$ the social loss for the taxi, if τ_{xy}^{γ} is the time waited by the taxi. The loss functions l_{xy} should verify $l_{xy}\left(0\right) = 0$ and that $l_{xy}\left(t\right) > 0$ for $t > 0$.

If $(\mu, \tau^{\alpha}, \tau^{\gamma})$ is the equilibrium outcome under rationing-by-waiting associated to systematic utilities α and γ, and marginal distributions of types n and m, then the total welfare loss is

$$L\left(\alpha, \gamma, n, m\right) = \sum_{xy} \mu_{xy} \left(l_{xy}^{G}\left(\tau_{xy}^{\alpha}\right) + l_{xy}^{H}\left(\tau_{xy}^{\gamma}\right)\right).$$

As one could expect, the total welfare loss is zero if and only if the market is equivalent to a market with transfers.

Proposition 5. *Assume that the distributions $\mathbf{P}_x$ and $\mathbf{Q}_y$ have a nonvanishing density. Let $(\tau^{\alpha}, \tau^{\gamma})$ be the equilibrium outcome 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*

$$\max_{\mu \geq 0} \left\{ \sum_{xy} \mu_{xy} \left(\alpha_{xy} + \gamma_{xy}\right) - \mathcal{E}\left(\mu\right) \right\} \tag{5.14}$$

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.

Proof. Because of the assumption made on $\mathbf{P}_x$ and $\mathbf{Q}_y$, $\mu_{xy} > 0$ for every x and y. Thus $L\left(\alpha, \gamma, n, m\right) = 0$ if and only if $\tau_{xy}^{\alpha} = 0$ and $\tau_{xy}^{\gamma} = 0$ for every x and y. Hence, if $L\left(\alpha, \gamma, n, m\right) = 0$, then $\mu = \nabla G\left(\alpha\right) = \nabla H\left(\gamma\right)$, which implies $\alpha + \gamma = \nabla G^{*}\left(\mu\right) + \nabla H^{*}\left(\mu\right) = \nabla \mathcal{E}\left(\mu\right)$, which is the first order condition associated to the strictly concave

maximization problem (5.14). Conversely, if μ is a solution of (5.14), 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 $L(\alpha,\gamma,n,m) = 0$. ∎

Formulation (5.14), which was introduced in in Galichon and Salanié (2015), characterizes matching models with transferable utility. We note that when $L(\alpha,\gamma,n,m) = 0$, 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).

We define exponential welfare loss functions as $l(t) = \exp(t) - 1$. It turns out that in the case when all the losses functions are exponential, we get:

Example 2 (continued). *In the logit case, and when the welfare loss functions l^G_{xy} and l^H_{xy} are all exponential, the total welfare loss can expressed as*

$$\begin{aligned} L(\alpha,\gamma,n,m) &= \sum_{x\in\mathcal{X}}\mu_{x0}\left(1+\sum_{y\in\mathcal{Y}}e^{\alpha_{xy}}\right)+\sum_{y\in\mathcal{Y}}\mu_{0y}\left(1+\sum_{x\in\mathcal{X}}e^{\gamma_{xy}}\right)-\sum_{x\in\mathcal{X}}n_x-\sum_{y\in\mathcal{Y}}m_y \\ &= \sum_{xy}\mu_{x0}\left(e^{\alpha_{xy}}-e^{U_{xy}}\right)+\mu_{0y}\left(e^{\gamma_{xy}}-e^{V_{xy}}\right). \end{aligned}$$

Indeed, it follows from the definition that $L(\alpha,\gamma,n,m) = \sum_{xy}\mu_{xy}\left(e^{\tau^\alpha_{xy}}+e^{\tau^\gamma_{xy}}-2\right)$, hence

$$\begin{aligned} L(\alpha,\gamma,n,m) &= \sum_{xy}\mu_{x0}e^{\alpha_{xy}-\tau^\alpha_{xy}}e^{\tau^\alpha_{xy}}+\sum_{xy}\mu_{0y}e^{\gamma_{xy}-\tau^\gamma_{xy}}e^{\tau^\gamma_{xy}}-2\sum_{xy}\mu_{xy} \\ &= \sum_{x\in\mathcal{X}}\mu_{x0}\sum_{y\in\mathcal{Y}}e^{\alpha_{xy}}+\sum_{y\in\mathcal{Y}}\mu_{0y}\sum_{x\in\mathcal{X}}e^{\gamma_{xy}}-\sum_{x\in\mathcal{X}}(n_x-\mu_{x0})-\sum_{y\in\mathcal{Y}}(m_y-\mu_{0y}) \\ &= \sum_{x\in\mathcal{X}}\mu_{x0}\left(1+\sum_{y\in\mathcal{Y}}e^{\alpha_{xy}}\right)+\sum_{y\in\mathcal{Y}}\mu_{0y}\left(1+\sum_{x\in\mathcal{X}}e^{\gamma_{xy}}\right)-\sum_{x\in\mathcal{X}}n_x-\sum_{y\in\mathcal{Y}}m_y, \end{aligned}$$

QED.

In the particular case of the logit model with exponential losses and when there is only one type on each side of the market, one can provide an expression for L in closed-from:

Example 3. *In the binomial case ($|\mathcal{X}| = |\mathcal{Y}| = 1$), one has*

$$\mu = \min\left(\frac{ne^{\alpha}}{1+e^{\alpha}}, \frac{me^{\gamma}}{1+e^{\gamma}}\right)$$

and

$$L(\alpha, \gamma, n, m) = ne^{\alpha} + me^{\gamma} - (2 + e^{\alpha} + e^{\gamma}) \min\left(\frac{ne^{\alpha}}{1+e^{\alpha}}, \frac{me^{\gamma}}{1+e^{\gamma}}\right).$$

5.4. **Surge pricing in matching markets.** In the surge pricing problem, the central planner sets a *surge price* of a unit of the service provided to x by y, denoted p_{xy}. It is assumed that the utilities are such that $\alpha_{xy}(p_{xy})$ is decreasing and continuous and $\gamma_{xy}(p_{xy})$ is increasing and continuous. The central planner seeks to chose the price vector (p_{xy}) in order to minimize the total market inefficiency. For the sake of exposition, we shall first review the benchmark case when there is no uncertainty on demand; in that case, under minimal additional assumptions, it is possible to set a price vector (p_{xy}) such that the total market inefficiency is zero, that is, market clears.

Theorem 5. *Assume that:*

(i) the distributions $\mathbf{P}_x$ and $\mathbf{Q}_y$ have a nonvanishing density,

(ii) the maps $p_{xy} \to \alpha_{xy}(p_{xy})$ are decreasing and continuous from $\mathbb{R}$ onto $\mathbb{R}$, and

(iii) the maps $p_{xy} \to \gamma_{xy}(p_{xy})$ are increasing and continuous from $\mathbb{R}$ onto $\mathbb{R}$.

Then there is a surge price vector $p = (p_{xy})$ such that the total market inefficiency is zero, that is

$$L(\alpha(p), \gamma(p), n, m) = 0. \tag{5.15}$$

Proof. The proof is short as it essentially consists of reformulating problem (5.15) as a matching problem with imperfectly transferable utility and heterogeneity in preferences as in Galichon et al. (2016). Clearly, $L(\alpha(p), \gamma(p), n, m) = 0$ if and only if $\nabla G(\alpha(p)) = \nabla H(\gamma(p))$. Letting $\mathcal{F}_{xy} = \left\{(u, v) \in \mathbb{R}^2 : \exists p, u \leq \alpha(p) \text{ and } v \leq \gamma(p)\right\}$, the models satisfies the assumptions of theorem 1 in that paper, and as a result, there exists p such that $\nabla G(\alpha(p)) - \nabla H(\gamma(p)) = 0$, which implies that (5.15) holds. ∎

Theorem 5 is benchmark will full information, but it would be interesting to investigates what happens when the central planner who sets the prices has limited availability on supply and demand. This will be the case, for instance, of an online matching platform deciding in real-time on the surge prices with imperfect forecast of supply and demand. Hence, we now discuss the case when the demand and supply vectors $\tilde{n}$ and $\tilde{m}$ are random. Then, the central planner needs to set p in order to minimize the expected market inefficiency, that is

$$\min_p \mathbb{E}\left[L\left(\alpha\left(p\right),\gamma\left(p\right),\tilde{n},\tilde{m}\right)\right].$$

In the binomial case of example 3, recall that we have a closed-form expression for the loss function

$$L\left(\alpha,\gamma,n,m\right)=ne^{\alpha}+me^{\gamma}-\left(2+e^{\alpha}+e^{\gamma}\right)\min\left(\frac{ne^{\alpha}}{1+e^{\alpha}},\frac{me^{\gamma}}{1+e^{\gamma}}\right).$$

Assuming that the demand $\tilde{n}$ is stochastic and lognormal, with expression $\tilde{n}=\bar{n}e^{\sigma g-\sigma^2/2}$ where $g\sim N\left(0,1\right)$, and that the supply vector m is nonstochastic, we have

$$\begin{aligned}\mathbb{E}\left[L\left(\alpha,\gamma,\tilde{n},m\right)\right] &= \bar{n}e^{\alpha}+me^{\gamma}-\left(2+e^{\alpha}+e^{\gamma}\right)\mathbb{E}\left[\min\left(\frac{\bar{n}e^{\alpha}}{1+e^{\alpha}}e^{\sigma g-\sigma^2/2},\frac{me^{\gamma}}{1+e^{\gamma}}\right)\right]\\ &= \bar{n}e^{\alpha}+me^{\gamma}-\left(2+e^{\alpha}+e^{\gamma}\right)P_{\sigma}\left(\frac{\bar{n}e^{\alpha}}{1+e^{\alpha}},\frac{me^{\gamma}}{1+e^{\gamma}}\right)\end{aligned}$$

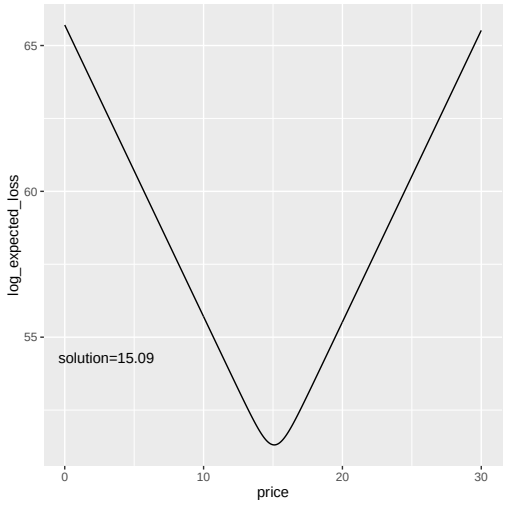
where $P_{\sigma}\left(S,K\right)$ is the value of the put in Black and Scholes' formula, given by

$$P_{\sigma}\left(S,K\right)=KN\left(-\frac{\ln\left(S/K\right)+\sigma^2/2}{\sigma}\right)-SN\left(-\frac{\ln\left(S/K\right)-\sigma^2/2}{\sigma}\right)$$

Assume $\alpha\left(p\right)=\alpha_0-p$ and $\gamma\left(p\right)=\gamma_0+p$, then the problem boils down to

$$\min_p\left\{\bar{n}e^{\alpha_0-p}+me^{\gamma_0+p}-\left(2+e^{\alpha_0-p}+e^{\gamma_0+p}\right)P_{\sigma}\left(\frac{\bar{n}e^{\alpha_0-p}}{1+e^{\alpha_0-p}},\frac{me^{\gamma_0+p}}{1+e^{\gamma_0+p}}\right)\right\}.$$

FIGURE 1. Expected Loss Function (log scale)



Parameter setup: $\alpha_0 = 60$, $\gamma_0 = 30$, $\bar{n} = 150$, $m = 100$, $\sigma = 10$.

Figure 1 illustrates this result by plotting the social loss as a function of the surge price.

6. Link with classical theory

In this final section, we investigate what happens when there is no stochastic utility component, i.e. in the case when $\varepsilon_y = 0$ for all $y \in \mathcal{Y}_0$ and $\eta_x = 0$ for all $x \in \mathcal{X}_0$. This case is not strictly covered by our previous framework, which assumed that the distributions of ε and η have a nonvanishing density. In this section we suitably adapt the definition of stable matching with rationing-by-waiting in the case without stochastic utility component.

6.1. **Stable matching with rationing-by-waiting with fully determinstic utility.** We consider a version of the previous model without a stochastic utility component. As before, there are $n_x \in \mathbb{N}$ taxis of type $x \in \mathcal{X}$ and $m_y \in \mathbb{N}$ passengers of type $y \in \mathcal{Y}$, and the surplus obtained by passenger x riding in a taxi of type y after waiting for τ^{α}_{xy} is $\alpha_{xy} - \tau^{\alpha}_{xy}$, while the surplus for taxi y of transporting a passenger of type x after waiting for τ^{γ}_{xy} is $\gamma_{xy} - \tau^{\gamma}_{xy}$. As before, μ_{xy} is the number of passengers of type x riding in taxis of type y. The reservation utilities of passengers and taxis is normalized to zero without loss of generality.

The stability condition expresses the fact that there cannot be simultaneously a nonempty waiting line of taxis and of passengers in the market segment xy, that is

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

At equilibrium, passengers choose the taxi bringing them the maximum amount of utility; similarly, taxis choose their utility-optimal passenger. Letting u_x and v_y be the indirect utilities of passengers of type x and taxis of type y, respectively, one has

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

and thus $u_x - \alpha_{xy} \geq -\tau_{xy}^{\alpha}$ with equality if x chooses y, that is if $\mu_{xy} > 0$. Similarly, $v_y - \gamma_{xy} \geq -\tau_{xy}^{\gamma}$ with equality if $\mu_{xy} > 0$. Hence,

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

Now, if $\mu_{xy} > 0$, then $u_x - \alpha_{xy} = -\tau_{xy}^{\alpha}$ and $v_y - \gamma_{xy} = -\tau_{xy}^{\gamma}$, and hence $\max\left(u_x - \alpha_{xy}, v_y - \gamma_{xy}\right) = \max\left(-\tau_{xy}^{\alpha}, -\tau_{xy}^{\gamma}\right) = 0$.

This brings us to the following definition of a stable outcome in the matching problem with deterministic utilities:

Definition 2. *An outcome (μ, u, v) is a stable matching with rationing-by-waiting with deterministic utilities 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* $\mu_{0y} := m_y - \sum_{x\in\mathcal{X}} \mu_{xy} > 0$.

A series of remarks are in order.

Remark 6.1. Clearly, this definition offers a close parallel with the definition of stable matching with transferable utility of Becker (1973) and Shapley-Shubik (1972). By simply replacing the max function by the sum in point (iv), definition 2 would become the definition of stable matching with transferable utility.

Remark 6.2. Definition 2 is not equivalent to the classical definition of stable matchings with non-transferable utility. However, it is closely connected, and equivalent in an important case, as argued in section 6.2 below.

Remark 6.3. As explained in section 6.3 below, the stable matchings with deterministic utilities as introduced in definition 2 may be obtained as limits of the stable matchings with stochastic utility introduced in definition 1.

6.2. **Comparison with classical stable matchings.** We would like to compare this notion with the classical notion of stable matching introduced in Gale and Shapley (1962). We have seen in example 1 that when there are several indistinguishable individuals per type, the notions of stable matching with nonprice rationing may depart from the classical one. However, when there is only one individual of each type, we shall see that the two notions are equivalent in some sense.

In this paragraph, we shall assume that $n_x = 1$ for all $x \in \mathcal{X}$ and $m_y = 1$ for all $y \in \mathcal{Y}$. In this case, μ is a stable matching in the classical sense if and only if:

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

(ii) $\sum_{y \in \mathcal{Y}} \mu_{xy} \leq 1$

(iii) $\sum_{x \in \mathcal{X}} \mu_{xy} \leq 1$

(iv) there is no blocking pair xy such that $\alpha_{xy} > \overline{u}_x^\mu$ and $\gamma_{xy} > \overline{v}_y^\mu$ hold together, where

$$\overline{u}_x^\mu := \sum_{y' \in \mathcal{Y}} \mu_{xy'} \alpha_{xy'} \text{ and } \overline{v}_y^\mu := \sum_{x' \in \mathcal{X}} \mu_{x'y} \gamma_{x'y}$$

(v) for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, $\sum_{y' \in \mathcal{Y}} \mu_{xy'} \alpha_{xy'} \geq 0$ and $\sum_{x' \in \mathcal{X}} \mu_{x'y} \alpha_{x'y} \geq 0$.

The following result shows that when there is only one individual per type, the notions of stable matching with rationing-by-waiting is equivalent to the notion of stable matching in the classical sense.

Theorem 6. *Let $n_x = 1$ for all $x \in \mathcal{X}$ and $m_y = 1$ for all $y \in \mathcal{Y}$. Assume μ is a stable matching in the classical sense. Then, setting $u_x = \sum_{y' \in \mathcal{Y}} \mu_{xy'} \alpha_{xy'}$ and $v_y = \sum_{x' \in \mathcal{X}} \mu_{x'y} \alpha_{x'y}$, it follows that (μ, u, v) is a stable matching with rationing-by-waiting.*

Conversely, assume (μ, u, v) is a stable matching with rationing-by waiting. Then μ is a stable matching in the classical sense, and $u_x \leq \overline{u}_x^\mu$ for all $x \in \mathcal{X}$ and $v_y \leq \overline{v}_y^\mu$ for all $y \in \mathcal{Y}$.

Proof. Assume μ is a stable matching in the classical sense. Then the stability inequality $\max\left(\overline{u}_x^\mu - \alpha_{xy}, \overline{v}_y^\mu - \gamma_{xy}\right) \geq 0$ holds, with equality if $\mu_{xy} = 1$; while $\overline{u}_x^\mu \geq 0$ with equality if $\mu_{x0} = 1$ and $\overline{v}_y^\mu \geq 0$ with equality if $\mu_{0y} = 0$. Thus, $(\mu, \overline{u}_x^\mu, \overline{v}_y^\mu)$ is a stable matching with rationing-by-waiting (without actual waiting time).

Conversely, consider (μ, u, v) a stable matching with rationing-by waiting, and let us show that μ is a stable matching in the classical sense. Assume (x, y) is a blocking pair such that $\alpha_{xy} > \overline{u}_x^\mu$ and $\gamma_{xy} > \overline{v}_y^\mu$, and let x' be the partner of y under μ and y' be the partner of x under μ, hence $\alpha_{xy} > \alpha_{xy'}$ and $\gamma_{xy} > \gamma_{x'y}$. One has $u_x - \alpha_{xy'} > u_x - \alpha_{xy}$ and $v_y - \gamma_{x'y} > v_y - \gamma_{xy}$, hence $\max\left(u_x - \alpha_{xy'}, v_y - \gamma_{x'y}\right) > \max\left(u_x - \alpha_{xy}, v_y - \gamma_{xy}\right) \geq 0$. Thus either $u_x > \alpha_{xy'}$ or $v_y > \gamma_{x'y}$; without loss of generality assume that $u_x > \alpha_{xy'}$. But because x and y' are matched under μ, it follows from the fact that (μ, u, v) a stable matching with rationing-by waiting that $\max\left(u_x - \alpha_{xy'}, v_{y'} - \gamma_{xy'}\right) = 0$, hence $u_x \leq \alpha_{xy'}$, a contradiction. ∎

Theorem 6 implies the following corollary, which is a mere restatement of its first assertion:

Corollary 1. *When there is one individual of each type, any stable matching in the classical sense can be interpreted as an equilibrium with rationing-by-waiting supported by zero waiting times.*

This result comes from the fact that with only one individual per type, the problem highlighted in example 1 vanishes. In that example, two identical individuals had different

prices, breaking down the existence of an equilibrium. Here, because there is only one individual per type, this type of pathology cannot occur.

6.3. **Limit when the stochastic utility component is small.** In this paragraph, we would like to show that stable matchings with rationing-by-waiting and logit stochastic component converg (when the amount of randomness tends to zero) to stable matchings with rationing-by-waiting and deterministic utilities. As a result, this will establish that such stable matchings exist, which is an not obvious fact per se. To do this, consider a model where the stochastic utility components are logit with scaling parameter $\sigma > 0$. The equilibrium matching μ_{xy} is given by

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

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

$$\left\{\begin{array}{l} \mu_{x0}(\sigma) + \sum_y \min\left(\mu_{x0}(\sigma)\, e^{\alpha_{xy}/\sigma}, \mu_{0y}(\sigma)\, e^{\gamma_{xy}/\sigma}\right) = n_x \\ \mu_{0y}(\sigma) + \sum_x \min\left(\mu_{x0}(\sigma)\, e^{\alpha_{xy}/\sigma}, \mu_{0y}(\sigma)\, e^{\gamma_{xy}/\sigma}\right) = m_y \end{array}\right\}.$$

Then, the following theorem holds:

Theorem 7. *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 a stable matching with rationing-by-waiting with deterministic utilities.*

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

Assume $\mu^*_{x0} > 0$. Then $-\sigma_n \ln \mu_{x0}(\sigma) \approx -\sigma_n \ln \mu^*_{x0} \to 0$ as $n \to +\infty$, thus $u_x = 0$.

Similarly, $\mu^*_{0y} > 0$ implies that $v_y = 0$. We have

$$\begin{aligned} \max\left(u_x - \alpha_{xy}, v_y - \gamma_{xy}\right) &= \lim_{n\to+\infty} \max\left(-\sigma_n \ln \mu_{x0}(\sigma_n) - \alpha_{xy}, -\sigma_n \ln \mu_{0y}(\sigma_n) - \gamma_{xy}\right) \\ &= \lim_{n\to+\infty} \left\{-\sigma_n \ln\left(\mu_{xy}(\sigma_n)\right)\right\} \end{aligned} \tag{6.1}$$

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

$$\max\left(u_x - \alpha_{xy}, v_y - \gamma_{xy}\right) \geq 0. \tag{6.2}$$

Assume $\mu^*_{xy} > 0$. Then $\sigma_n \ln\left(\mu_{xy}(\sigma_n)\right) \to 0$ as $n \to +\infty$, and therefore inequality (6.2) is saturated. Thus (μ^*, u, v) satisfy conditions (ii) to (vi) of definition 2, 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 2. 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 $\mu^*_{xy} > 0$. QED. ∎

As a corollary, we get that there always exists an equilibrium with rationing-by-waiting with deterministic utilities. This existence followed from theorem 6 in the case when there is one number of individuals per type, but the following result shows that this is true regardless of the number of individuals of each types.

Corollary 2. *There always exists an equilibrium with rationing-by-waiting.*

§ *New York University, Department of Economics and Courant Institute, 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, 3620 South Vermont Ave. Kaprielian Hall Suite 300, Los Angeles, CA 90089, USA. Email: yuwei.hsieh@usc.edu*

References

[1] Arnott, R. (1996): Taxi Travel Should Be Subsidized, *Journal of Urban Economics*, 40(3), pp. 316-333.

[2] 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.

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

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

[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] Buchholz, N. (2016): Spatial Equilibrium, Search Frictions and Efficient Regulation in the Taxi Industry, working paper.

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

[8] Castillo, J. C., D. Knoepflez, and E. G. Weyl (2017): Surge Pricing Solves the Wild Goose Chase, working paper.

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

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

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

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

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

[14] 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.

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

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

[17] Frechette, G., A. Lizzeri, and T. Salz (2016): Frictions in a Competitive, Regulated Market: Evidence from Taxis, working paper.

[18] Fudenberg D., Iijima, R. and Strzalecki, T. (2015): Stochastic Choice and Revealed Perturbed Utility, *Econometrica*, 83(6), pp. 2371–2409

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

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

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

[22] Galichon, A. Kominers, S. and Weber, S. (2016): Costly Concessions: An Empirical Framework for Matching with Imperfectly Transferable Utility, working paper.

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

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

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

[29] Joskow, P., and Wolfram, C. (2012): Dynamic Pricing of Electricity, *American Economic Review*, 102(3), pp. 381-385.

[30] Levin, J. and Skrzypacz, A. (2016): Platform Pricing for Ride-Sharing. Working paper.

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

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

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

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

[35] McAfee, R. P. and V. t. Velde (2006): Dynamic Pricing in the Airline Industry, *Handbook on Economics and Information Systems*, Chapter 11, pp. 527–567, Ed: TJ Hendershott, Elsevie.

[36] 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.

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

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

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

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

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

[42] 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.

[43] 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.

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

[45] 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.

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

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

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

[49] Williamson, O. (1966). "Peak-Load Pricing and Optimal Capacity under Indivisibility Constraints." *American Economic Review* 56, No. 4, Part 1, pp. 810-827.