Matching and Network Effects in Ride-Hailing

By Juan Camillo Castillo and Shreya Mathur

Online Appendix

Full model and proofs of results

A1. Supply, demand, equilibrium, and welfare

Suppose that the arrival rate of passengers is given by a demand function

$$Q = D(p, T_B) \tag{A1}$$

that is decreasing in both p, the price paid by buyers, and T_B, the average waiting time buyers need to wait before being matched.

The number of sellers is determined by a decreasing labor supply curve

$$N = L(w), \tag{A2}$$

where w are sellers' average earnings per-unit-time. To find an expression for w, let p' be the amount sellers get per transaction, which can be different from p if there is a per-unit tax or subsidy or if matching occurs through a platform that gets a commission. Additionally, let c be a marginal cost per transaction. Sellers' average earnings are then given by $w = \frac{(p'-c)Q}{N}$.

As stated in the main text, the number of matches per unit time are given by the matching function $m(B, S)$. We assume that it is continuous and increasing in both arguments, and that $m(0, B) = m(B, 0) = 0$. It may have a Cobb-Douglas function $m(B, S) = MB^{\alpha}S^{\beta}$. Lagos (2003), for instance, microfounds a Cobb-Douglas matching function for taxis with $\alpha = \beta = 1$ based on a spatial search model.

The steady state is defined by equation (1). The following result guarantees existence and uniqueness of a steady state under certain conditions:

PROPOSITION 1: *If $tQ < N$ and $\lim_{B\to\infty} m(B, N - tQ) \geq Q$, then there exists a unique (B, S) that is consistent with steady state. Otherwise, there is no steady state.*

PROOF:

Conside (Q, N) such that $tQ < N$. S is pinned down by the equation on the number of drivers, $S = N - tQ$. B is defined implicitly by $Q = m(B, S)$. If $\lim_{B\to\infty} m(B, N - tQ) \geq Q$, it has a unique solution by the intermediate value theroem since $m(0, S) = 0$, $m(B', S) > Q$ for some B', and $m(B, S)$ is increasing in B.

If $tQ > N$, there is no nonnegative S that satisfies the identity for drivers $N = S + tQ$. If $\lim_{B\to\infty} m(B, N - tQ) < Q$, then there exists no solution to $Q = m(B, S) = m(B, N - tQ)$ because $m(B, S)$ is increasing in B. Note that $tQ = N$ implies $S = 0$, so $\lim_{B\to\infty} m(B, N - tQ) = 0 < Q$. ■

This proposition means that when there are too many buyers there are not enough sellers to serve them. But as long as that is not the case, an equilibrium exists and is unique.

The steady-state waiting time for buyers is given by $\tilde{T}_B(Q,N)$. A market equilibrium is a solution in Q and N of

$$Q = D(p, \tilde{T}_B(Q,N)) \qquad N = L\left(\frac{(p'-c)Q}{N}\right). \tag{A3}$$

There may be multiple equilibria given p and p'. However, for any (Q,N) there exists a unique set $(p(Q,N), p'(Q,N))$ that is consistent with market equilibrium.[3] Thus, although Q and N are endogenous, it will be more convenient to work with quantities instead of prices.[4]

Welfare is the sum of buyer surplus, seller surplus, and the net payments obtained by the platform setting prices or the government charging taxes or paying subsidies:

$$W = BS + SS + (p - p')Q, \tag{A4}$$

where we define buyer surplus and seller surplus as

$$BS(p, T_B) = \int_p^\infty D(\tilde{p}, T_B)\, d\tilde{p} \qquad SS(w) = \int_0^w L(\tilde{w})\, d\tilde{w}. \tag{A5}$$

We also define buyers' gross utility and sellers' labor cost as their surplus without netting out their payments, i.e.,

$$U(p, T_B) = BS(p, T_B) + pD(p, T_B) \qquad -C(w) = SS(w) - wL(w). \tag{A6}$$

If we define inverse demand and supply functions $p(Q, T_B)$ and $w(N)$ (which we can do since supply and demand are decreasing in prices and increasing in expected earnings, respectively), we can define gross utility and supply cost in terms of quantities in order to write welfare as in equation (2).[5]

By using a change of variables, we can rewrite gross utility and labor cost as integrals over quantities:[6]

$$U(Q, T_B) = \int_0^Q p(\tilde{Q}, T_B)\, d\tilde{Q} \qquad C(N) = \int_0^N w(\tilde{N})\, d\tilde{N} \tag{A7}$$

[3] To see that, define $p(Q,N)$ and $p'(Q,N)$ as the implicit solutions to (A3). Both are well-defined because demand is decreasing in p and supply is decreasing.

[4] The only concern is that the market could end up in a wrong equilibrium $(Q',N') \neq (Q,N)$. However, a price-setting platform can ensure the right equilibrium plays out in the long run: if agents expect an equilibrium (Q',N'), it can set prices $p(Q',N')$ and $p'(Q',N')$ until agents' expectations adjust. It can then revert to the equilibrium prices for the desired allocation. This follows the idea of an "insulating tariff" from Weyl (2010).

[5] The algrebra, without explicitly writing function arguments, is $W = BS + SS - pQ - p'Q = U - pQ + wN - C + pQ - p'Q = U + (p' - c)Q - C - p'Q = U - C - cQ$.

[6] This is most easily seen by plotting the supply and demand curves. Gross utility is the are below the demand curve and the labor cost is the area below the labor supply curve.

From these expressions, it is clear that $\frac{\partial U}{\partial Q} = p$ and $\frac{\partial C}{\partial N} = w$.

A2. Returns to scale

We start by defining formally what we mean by returns to scale.

DEFINITION 1: *The matching technology has*

- Constant returns to scale *if* $m(aB, aS) = am(B, S)$ *for all* B, S, a.
- Increasing returns to scale *if* $m(aB, aS) > am(B, S)$ *for all* B, S *and for all* $a > 1$.
- Decreasing returns to scale *if* $m(aB, aS) < am(B, S)$ *for all* B, S *and for all* $a > 1$.

We now state three equivalent characterizations of increasing returns to scale.

PROPOSITION 2: *Each of the following statements holds if and only if the matching technology has increasing returns to scale:*

1) $T_B(aB, aS) < T_B(B, S)$ *for all* Q, N *and for all* $a > 1$.

2) $T_S(aB, aS) < T_S(B, S)$ *for all* Q, N *and for all* $a > 1$.

3) $\tilde{T}_B(aQ, aN) < \tilde{T}_B(Q, N)$ *for all* Q, N *and for all* $a > 1$.

PROOF:

We prove the three statements one by one.

1) *Setup:* Take any B and S, and take any $a > 1$.

   If: Suppose that $m(aB, aS) > am(B, S)$. By the definition of T_B, $T_B(aB, aS) = \frac{aB}{m(aB,aS)} < \frac{aB}{am(B,S)} = \frac{B}{m(B,S)} = T_B(B, S)$.

   Only if: Suppose that $T_B(aB, aS) < T_B(B, S)$. By the definition of T_B, $m(aB, aS) = \frac{aB}{T_B(aB,aS)} > \frac{B}{aT_B(B,S)} = \frac{B}{T_B(B,S)} = m(B, S)$.

2) The proof is entirely analogous to the previous part, substituting S for B whenever necessary.

3) *Setup:* Let B and S be the equilibrium quantities with (Q, N). Now take another equilibrium with (aQ, aN) where $a > 1$. By the balance equation for sellers, the number of searching sellers is $aN - taQ = aS$. Let the number of searching buyers in that other equilibrium be bB. Then $\tilde{T}_B(aQ, aN) = \frac{bB}{m(bB,aS)} = \frac{bB}{aQ}$.

   If: Suppose that $m(aB, aS) > am(B, S)$. By the balance equation for buyers, $m(bB, aS) = aQ$. By the increasing returns to scale, $m(aB, aS) > aQ = m(bB, aS)$, which implies that $a > b$ since $m(B, S)$ is increasing in B. Finally, note that $\tilde{T}_B(aQ, aN) = \frac{bB}{m(bB,aS)} = \frac{bB}{aQ} = \frac{b}{a}\frac{B}{Q} = \frac{b}{a}\tilde{T}_B(Q, N) < \tilde{T}_B(Q, N)$.

   Only if: Suppose that $\tilde{T}_B(aQ, aN) < \tilde{T}_B(Q, N)$. This can be rewritten as $\frac{B}{Q} > \frac{bB}{m(bB,aS)}$, which implies that $a > b$. Thus, $m(aB, aS) > m(bB, aS) = am(B, S)$, where the inequality arises from the fact that $m(B, S)$ is increasing in b. ■

Proposition 2 only states results for increasing returns to scale, but analogous results also hold for decreasing and constant returns to scale. The proofs follow the exact same process, flipping inequalities or changing them to equalities. One may expect that a result resembling part 3 also holds for $\tilde{T}_S$, but that is not the case. In fact, $\tilde{T}_S$ is always homogeneous of degree zero.[7]

A3. Welfare maximization

The two equations for welfare maximization (equating (3) and (4) to zero), can be written as $p = c - \frac{\partial U}{\partial T_B}\frac{\partial T_B}{\partial Q}$ and $\frac{(p'-c)Q}{N} = \frac{\partial U}{\partial T_B}\frac{\partial T_B}{\partial N}$. From these two expressions, $p - p' = \bar{\tau} = -\frac{\partial U}{\partial T_B}\left(\frac{\partial T_B}{\partial Q} + \frac{N}{Q}\frac{\partial T_B}{\partial N}\right) = -(\epsilon_Q^{T_B} + \epsilon_N^{T_B})T\bar{u}_T$.

This last equation illustrates the relationship between returns to scale and taxes/subsidies, but it is not amenable for the computation because $\epsilon_Q^{T_B}$ and $\epsilon_N^{T_B}$ are properties of market equilibria and not of primitives of the model. An equivalent expression that we use for our main results is $\bar{\tau} = -\frac{\epsilon_S^{T_B} + \epsilon_B^{T_B}}{1-\epsilon_B^{T_B}} T\bar{u}_T$, which only depends on properties of the matching function.[8]

Computing the matching function for ride-hailing

B1. Matching in ride-hailing

Let A be the density of available drivers. The pickup time is a decreasing, convex function $P(A)$. As an example, suppse that drivers move at a constant speed in a homogeneous n-dimensional space. In that case, $P(A) \propto \frac{1}{A^{\frac{1}{n}}}$.

Assuming riders are matched immediately to their nearest driver, they only need to wait while they are picked up, so that $T_B = P(A)$. The driver first needs to wait for a period of time $\frac{A}{m}$ while they are matched, and then they need to pick up the passenger, so that the total pickup time is equal to $T_S = \frac{A}{m} + P(A)$. The waiting times of riders and drivers give a system of two equations on two unknowns (m and A):

$$T_R = P(A) \qquad T_D = \frac{A}{m} + P(A). \tag{B1}$$

[7] Let B and S be the equilibrium quantities with (Q, N). Now take another equilibrium with (aQ, aN). The number of waiting sellers is $aN - taQ = aS$. So $\tilde{T}_S(aQ, aN) = \frac{aS}{aQ} = \frac{S}{Q} = \tilde{T}_S(Q, N)$.

[8] To derive this expression, note first that, since $T_B = \frac{B}{Q}$, we can write $d\log T_B = \epsilon_N^B \cdot d\log N + (\epsilon_Q^B - 1)\cdot d\log Q$, which implies that $\epsilon_Q^{T_B} + \epsilon_N^{T_B} = \epsilon_Q^B + \epsilon_N^B - 1$. And from $Q = m(B, N - tQ)$ (which combines both steady-state conditions), we get $d\log Q = \epsilon_B^m \cdot d\log B + \epsilon_S^m \cdot (\frac{N}{S}\cdot d\log N - \frac{tQ}{S}\cdot d\log Q)$. Isolating $d\log B$ yields $d\log B = \frac{1}{\epsilon_B^m}\left[-\epsilon_B^m \frac{N}{S} d\log N + (1 + \epsilon_S^m \frac{tQ}{S}) d\log Q\right]$. From that we can derive $\epsilon_N^B + \epsilon_Q^B = \frac{1}{\epsilon_B^m}\left[1 + \epsilon_S^m\left(\frac{tQ}{S} - \frac{N}{S}\right)\right]$, and if we note that $tQ - N = -S$, that last expression is $\epsilon_N^B + \epsilon_Q^B = \frac{1-\epsilon_S^m}{\epsilon_B^m}$.

Combining both expressions we have derived for elasticities, we obtain $\epsilon_Q^{T_B} + \epsilon_N^{T_B} = \frac{\epsilon_S^{T_B} + \epsilon_B^{T_B}}{1-\epsilon_B^{T_B}}$.

Solving for m and substituting $T_B = \frac{B}{m}$ and $T_S = \frac{S}{m}$ gives us the matching function:

$$m(B,S) = \frac{B}{P(S-B)}. \tag{B2}$$

This function exhibits increasing returns to scale because $P(A)$ is a decreasing function: $m(aB,aS) = \frac{aB}{P(aS-aB)} > \frac{aB}{P(S-B)} = am(B,S)$.

As a simple example, suppose that $P(A) \propto \frac{1}{A^{-\gamma}}$. That is the case, for instance, in the homogeneous n-dimensional space discussed above, with $\gamma = \frac{1}{n}$. In that case, $m(B,S)$ is homogeneous of degree $1+\gamma > 1$, and, hence, matching has IRS. If space is homogeneous and two-dimensional, and drivers move at a constant speed, $\gamma = 0.5$. However, drivers do not drive at constant speed. As shown by Castillo et al. (2022), they drive slower on short trips due to the structure of roads. The pickup times of short trips thus tend to be somewhat higher than if $P(A) \propto \frac{1}{A^{-\gamma}}$. In other words, the pickup times tend to be somewhat higher if A is high, as would be the case if $\gamma < 0.5$. This theoretically justifies a matching function that is homogeneous of degree ~ 1.35.

We now highlight two important features about the matching function (B2). First, this function is only defined for $S > B$. Equation (B1) means that drivers always need to wait longer than drivers. As such, the number of waiting drivers is always higher than the number of waiting riders. Note also that the matching function is not necessarily increasing in B. That is the case whenever the market is under a matching failure that Castillo et al. (2022) call a "wild-goose chase." Many of our results cease to hold when that is the case. However, those are pathological conditions that we do not consider in this paper.

B2. Simulation

Suppose a rider requests a trip at a time t from coordinates x. Drivers are uniformly distributed around her with a probability density A.

Consider one draw from that density of drivers. It is characterized by (a) a set $\mathcal{J}$ of available drivers and (b) the coordinates y_j of each driver $j \in \mathcal{J}$. Matches take place as follows. First, the platform computes a pickup time w_j for every rider j. It is drawn from a distribution $G(\cdot|x,y_j,t)$ that depends on the coordinates of both the rider and the driver and on the time of the week when the trip was requested.

The platform first offers the trip to the driver in J_t with the lowest pickup time, who accepts it with probability ϕ. If he does not accept it, the trip is offered to the next closest driver, who also accepts it with probability ϕ. This process goes on until one driver j^* accepts the trip. The realized pickup time for the rider is w_{j*}.

Based on this process, we can compute $P(A,x,t)$, the pickup time function from the perspective of a rider requesting a trip from x at time t. We simply simulate the above process given (A,x,t), and we average the realized pickup times.

B3. Estimation

We need to estimate two elements to be able to compute $P(A, x, t)$: the density of pickup times $G(\cdot|x, y_j, t)$ and drivers' acceptance probability ϕ. We estimate them using data from all Uber trips as well as all trip offers to drivers in Houston between March 16 and April 8, 2017. We observe timestamps and coordinates of the rider and driver at three points in time for every trip: when the trip was requested, when the rider was picked up, and when the rider was dropped off. We also observe data on all trip offers to drivers and whether they were accepted.

To estimate pickup times, we first fit a random forest to obtain a prediction $\hat{T}(x, y_j, h)$ of pickup times as a function of rider and driver coordinates and the hour of the week. We also fit a linear model of the standard deviation of the residual of this model as a function of the prediction $\hat{T}$. Let $\hat{sd}(\hat{T})$ be the prediction from this model. To generate draws from $G(\cdot|x, y_j, t)$, we take draws from a lognormal distribution with mean $\hat{T}(x, y_j, h)$ and standard deviation $\hat{sd}(\hat{T}(x, y_j, h))$.

We take ϕ to be the average trip acceptance probability in the data, 0.847.

B4. Results

We focus on 36 markets, defined as combinations of six different locations and six hours of the week, which we believe roughly represent different types of neighborhoods in Houston. The six locations are Downtown Houston, Brays Oak, a middle-income largely residential neighborhood, the Third Ward, which contains the largest African-American population of Houston, University of Houston, the largest university in the city, Greater Heights, a dense neighborhood with a significant young professional population, and the Memorial area, a large commercial neighborhood surrounded by high-income residential neighborhoods. For each location, we take the latitude and longitude of one particular point. For Downtown Houston at 5 pm, for instance, we choose the point with coordinates (29.757629,-95.368672). We also focus on six times of the week: Tuesdays at 3 am, 8 am, 12 pm, and 5 pm, as well as Saturdays at 1 am and 3 pm.

Figure B1 shows a map with the predicted pickup times generated from the random forest for Downtown Houston on Tuesdays at 5 pm. As expected, the pickup duration is higher the further the driver is. And for a given distance, the pickup duration tends to be lower near highways, where drivers can drive quickly towards the rider.

Figure B2 plots the expected pickup time that we generate for different locations and times of the week. The pickup time is a decreasing function that is highest for congested markets such as during rush hour and Downtown. It is also high around the University of Houston, where many roads are closed to cars, making it harder for drivers to reach passengers.

In Figure B3 we plot the matching function for Downtown at Tuesday, 17 pm. The function is only defined for the upper left corner where the market is not under the matching failure that Castillo et al. (2022) call "wild-goose chases."

For our results about elasticities and externalities, we need to compute the distribution of waiting riders and drivers for every market. For Downtown Houston at

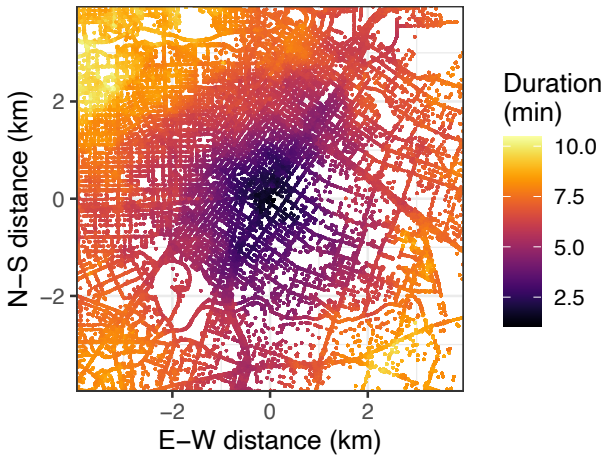
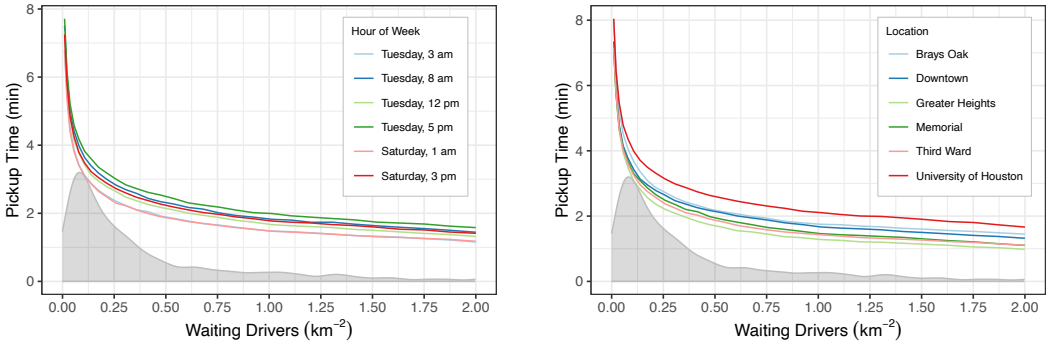


Figure B1. : Predicted Pickup Times

Note: This figure plots the predicted pickup times generated from the random forest for the downtown Houston area on Tuesdays at 5 pm. The points represent the locations where we observe available drivers in the data. For each driver location, we predict the pickup time to a passenger who requests a trip at the origin.



(a) Downtown

(b) Tuesday, 12 pm

Figure B2. : $P(A)$ - Pickup Times and Density of Drivers

Note: These figures plot the pickup time as a function of the driver density for the Downtown area at various times (subfigure a) and for different locations at Tuesday, 12 pm (subfigure b). The pickup times are computed using the pickup times generated from the random forest for different driver densities. The grey curve depicts the distribution of the driver densities that we observe in the data.

5 pm, for instance, we focus on everything that happens in a 4 km by 4 km square centered around the point with coordinates (29.757629,-95.368672) and between 4 pm and 6 pm on weekdays. We take the number of available drivers in the data and the number of drivers waiting to be picked up, both of which we aggregate into one hour periods. Each one of these periods constitutes one observation for the

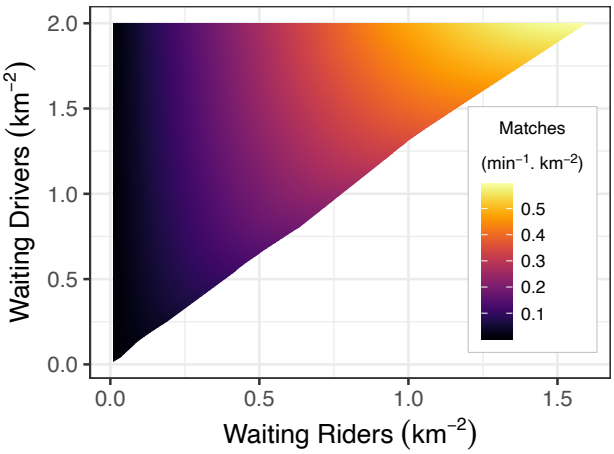
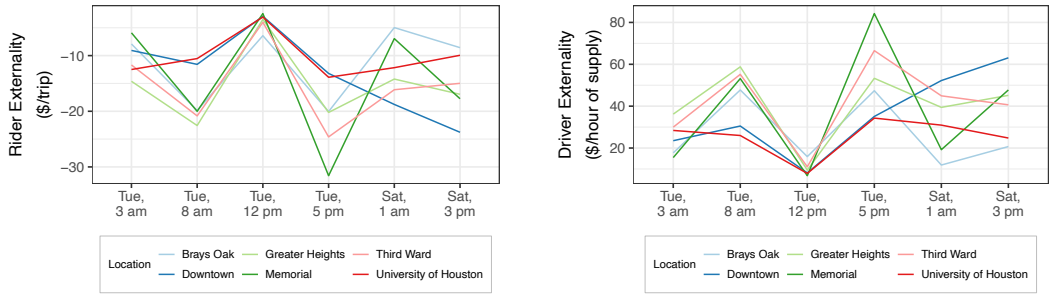


Figure B3. : Matching Function for Downtown at Tuesday, 5 pm

Note: This figure plots the matching function that we compute for Downtown Houston at 5 pm on Tuesdays. For each combination of the density of waiting riders and drivers, the function represents how many matches would occur per unit of time.

distribution of the density of riders and drivers.



(a) Rider Externality

(b) Driver Externality

Figure B4. : $P(A)$ - Pickup Times and Density of Drivers

Note: These figures plot the externalities riders and drivers, respectively, for different locations and hours of the week. For every market, the externalities that we show represent the average externality over the distribution of waiting riders and drivers that we observe in that market in the data.

Figure B4 plots the externalities caused by riders and drivers in dollars per trip and dollars per hour of work, respectively, for various locations and times. There is a much wider dispersion than we observe in the sum of both externalities, as we show in figure 2. The reason for this is that when the market is congested—when demand is high and there are relatively few drivers, such as during rush hour—the externalities caused by riders and drivers both become large in absolute terms. The

difference between them also becomes larger, but by not nearly as much as each individual externality.

*

Additional References

Lagos, Ricardo, "An Analysis of the Market for Taxicab Rides in New York City," *International Economic Review*, 2003, *44* (2), 423–434.

Weyl, E. Glen, "A Price Theory of Multi-Sided Platforms," *American Economic Review*, 2010, *100* (4), 1642–1672.