Spatial Distribution of Supply and the Role of Market Thickness: Theory and Evidence from Ride Sharing

Soheil Ghili and Vineet Kumar*

January 15, 2020

Abstract

This paper develops a strategy with simple implementation and limited data requirements to identify spatial distortion of supply from demand –or, equivalently, unequal access to supply among regions– in transportation markets. We apply our method to ride-level, multi-platform data from New York City (NYC) and show that for smaller rideshare platforms, supply tends to be disproportionately concentrated in more densely populated areas. We also develop a theoretical model to argue that a smaller platform size, all else being equal, distorts the supply of drivers toward more densely populated areas due to network effects. Motivated by this, we estimate a minimum required platform size to avoid geographical supply distortions, which informs the current policy debate in NYC around whether ridesharing platforms should be downsized. We find the minimum required size to be approximately 3.5M rides/month for NYC, implying that downsizing Lyft or Via–but not Uber–can increase geographical inequity.

JEL Codes: L13; R41; D62
Keywords: Spatial Markets; Transportation; Geographical Inequity; Market Thickness; Ridesharing

1 Introduction

In spatial markets, there are salient questions about possible geographical distortion of supply from demand: (i) how to empirically infer whether some regions are "under-supplied" relative to others; (ii) how to identify mechanisms that lead to unequal access to supply across regions; and (iii) how to design policies that alleviate geographical supply inequities.

This paper studies the above questions in the context of the ridesharing platforms Uber, Lyft, and Via in New York City (NYC henceforth). We provide a method with simple implementation

*Yale University. We thank Phil Haile, Igal Hendel, Larry Samuelson, and K. Sudhir for helpful comments.

and limited data requirements to detect spatial mismatch between supply and (potential) demand. We then offer empirical evidence along with a theoretical model that (to our knowledge, for the first time) points to the role of market "thickness" (i.e., platform size) in determining the spatial distribution of supply. We show that, all else being equal, in a thicker market, supply is geographically more balanced with demand. Finally, based on our theory results, we offer a simple empirical procedure to identify the "minimum required" rideshare platform size for a city to ensure geographical supply inequity will be negligible. We apply the method to NYC and find the minimum required size to be between 3.29M and 3.65M rides per month.

Mitigating inequity in access to supply of products/services among different regions of spatial markets can be a critical goal for policymakers. Indeed, it has led to such major actions as the launch of the Green Taxi in NYC to serve its outer boroughs. However, it is challenging to empirically measure the extent to which the arriving (potential) demand is more likely to go unfulfilled due to limited supply[1] in some regions, compared to other regions. The challenge arises from the fact that unfulfilled demand is usually unobserved. The empirical literature on spatial markets acknowledges this challenge. To address it, recent literature combines data on unmatched supply (e.g., empty cars) and matched demand/supply (e.g., realized rides) in each region with a structural model of matching, and then inverts the matching function in order to infer the level of unfulfilled demand. Variants of this method are being developed and leveraged by papers that study questions related to ours (Brancaccio et al., 2019a,b,c; Buchholz, 2018; Frechette et al., 2019).

The first portion of this paper proposes a method, called *relative outflows analysis*, that detects inequity among regions in the percent fulfillment of the unobserved arrival of potential demand[2] and apply it to the NYC rideshare market. Our approach is simple to implement as it does not require a model of matching. Additionally, it requires data on rides only (as opposed to rides *and* empty cars,) because our identification strategy leverages data not only on rides starting at any given region, but also on rides ending there. To illustrate, suppose Lyft's "relative outflow" in Staten Island (i.e., number of Lyft rides exiting Staten Island divided by number of those entering it) is persistently and substantially smaller than one, while Uber's is close to one. Assuming passengers are not using Lyft to permanently move to Staten Island, this can only mean the same population that chooses Lyft on its way into the region is on the average less likely to choose Lyft on its way out. Assuming that geographical heterogeneity of outside transportation options affects all rideshare platforms similarly, the persistently low relative outflow of Lyft in Staten Island cannot be attributed to attractive outside options in that region, given that Uber's relative outflow is high there. Rather, we interpret Lyft's small relative outflow to mean that potential demand for

[1] "Limited supply" can be in the form of high wait time and/or high price. See Section 4 for more details.

[2] Some papers in the literature think of potential demand as the number of those who search for rides. We think of it as those who search, plus those who decide not to search *only because the anticipate the search would fail.* See Section 4 for more details.

Lyft is going unfulfilled, disproportionately more in Staten Island than in the rest of the city, due to high wait times or prices. Section 4 details the implementation of our approach and lays out the assumptions under which a platform's relative outflow in a region can be interpreted as a (relative) measure of local access to supply for that platform. In addition to simplicity and limited data requirements, our method also has two other advantages. First, it can be readily applied to all passenger-transportation markets, no matter whether the matching system is centralized (e.g., rideshare) or decentralized (e.g., taxicabs). Second, if the supply of platform k in region i, relative to other regions, is so limited that passengers in i have learned not to search for rides with k, this regional under-supply gets reflected in k's relative outflow in i. This natural long-run reaction of passengers' search behavior would not be captured by models that aim only at inferring the number of passengers who searched but failed to find rides.[3]

The second component of our paper looks at why such geographical imbalance between supply and demand arises, and it studies the role of "market thickness." That is, we aim to study whether a smaller platform size can, all else being equal, skew the spatial distribution of supply more toward certain areas. We believe this paper is the first to examine the impact of market thickness on spatial distribution of supply. Papers that study spatial demand-supply mismatches (e.g., Buchholz (2018); Lagos (2000); Afèche et al. (2018); Banerjee et al. (2018)) focus on other mechanisms, mainly search frictions. Papers that study the consequences of market thickness (such as Frechette et al. (2019); Nikzad (2018)) have not looked at its implications for the geographical distribution of supply. We attempt to bring these two pieces together both empirically and theoretically.

On the empirical side, we start by documenting two data patterns. First, for each rideshare platform (Uber, Lyft, or Via,) the relative outflow is the largest in Manhattan and declines as we go toward outer, lower population density, boroughs. Second, the rate of decline is faster for smaller platforms (i.e., thinner markets). We formally test the latter pattern using a regression specification in which the dependent variable is relative outflow of a given platform in a given borough on a given date. The coefficient of interest is the interaction coefficient between the borough's population density and the platform's overall size (measured in rides/month across NYC). We show the estimated coefficient is robustly negative and significant across 72 combinations of functional form and fixed effects specifications. We interpret this robust result causally to mean that, *ceteris paribus*, a thinner market leads to increased under-supply in less densely populated areas.[4]

Our empirical analysis raises the question of whether there is any theoretical reason to expect market thickness (platform size) to influence the geographical distribution of supply. In Section 5, we develop a model of a monopolist rideshare platform that centrally matches drivers to riders,

[3] unless, as in some papers in the literature, that structural model is itself embedded within a model of passengers' decision making on whether to search for rides.

[4] More details on our interpretation of the results can be found in Section 4.2.1.

along with a fixed number of drivers who simultaneously decide in which of the $I \geq 2$ regions to operate. All regions have the same size but possibly different arrival rates of passengers (demand). Each driver chooses a region i that, given the choices of other drivers, will minimize his "total wait time." Total wait time consists of (i) "idle time," the time it takes for the driver to be assigned to a passenger requesting a ride, and (ii) "pickup time," the time it takes to arrive at the pickup location after being assigned to a passenger. More drivers in each region i means a higher expected idle time in i. This forces the supply of drivers to geographically distribute itself proportionally to the distribution of demand. On the other hand, more drivers in region i means a lower expected pickup time in i, forcing drivers to agglomerate. Our results study the interplay between these two forces.

We obtain three main results. First, the total number of drivers has to be large enough for there to exist an all-regions equilibrium–that is, one in which each region i gets a strictly positive number of drivers. Second, any all-regions equilibrium is unique and "excessively clustered" toward higher demand areas. That is, for any pair of regions, the equilibrium number of drivers divided by demand arrival is strictly larger in the region with higher demand. Finally, we study the impact of "thinning" the market either on one side only (a decrease in the total number of drivers) or on both sides (a proportional decrease in demand in each region and total number of drivers). We develop a inductive technique to prove that while each such thinning preserves the demand ratios, it skews the equilibrium supply ratio between *any* two regions toward the higher-demand region. The basic intuition is that the supply of drivers responds to a "global thinning" of the market, which increases pickup times everywhere, by further agglomerating in regions with "thicker local markets." Since these results are closely in line with our empirical findings, we believe our model provides a realistic understanding of the relationship between market thickness and spatial distribution of supply. The key deviation in our framework from the literature, which allows delivering our results, is that in our model of the spatial market, each region has a non-trivial "size" (rather than being a "point"). We use data on Uber's and Lyft's surge-price factors and estimated pickup times to provide suggestive evidence that what our model abstracts away from (mainly prices and platform competition) does not seem to play a first order role in leading to under-supply of rideshares in less dense areas. Additionally, we offer anecdotal evidence from rideshare forums that corroborates our theory and its relevance.

We close our theory section by discussing (i) the implications of geographical supply inequity for efficiency and (ii) the generality of our insights beyond ridesharing. On the efficiency front, we first show that by agglomerating in denser areas, each driver inefficiently increases the idle times of *other drivers* in the region he joins and the pickup times of other drivers in the regions he avoids. Such externalities lead to inefficiently high total wait times overall. Second, we argue that inequity among regions means persistent inequity among residents of those regions. This can be inefficient by making the marginal utility of rides heterogeneous across regions. We do not formally

model passengers' utility from rides, but we argue that this effect may have been the main motive for major policy acts such as the launch of Green Taxis. On the generality front, we argue that although pickup times play a crucial role in obtaining our theoretical results, similar phenomena are still likely to emerge in markets where matching is not centralized and, hence, pickup times are not present. We demonstrate this by observing that, in the Yellow Taxicab market, although pickup times are negligible (and instead, search frictions exist), the geographical pattern of relative outflows is qualitatively similar to, but quantitatively more pronounced than, that of rideshare: The relative outflows in higher population density boroughs are substantially higher than those in lower population density ones.

Last, Section 6 examines the policy implications of our results. Qualitatively, our analysis is complementary to some empirical and theoretical results in the literature (such as those in Frechette et al. (2019); Nikzad (2018)). These studies state that, though pro-competitive, breaking a large platform up might have some adverse effects for overall matching quality or service quality, because it may make the market for each (new and small) platform too thin. Our work suggests that breaking up a large platform may also decrease the geographical reach of supply whereby some areas are not served or are served poorly. On the quantitative side, we empirically estimate a critical rideshare platform size for NYC, above which the impact of size on the geographical distribution of relative outflows becomes negligible. This has two motivations. First, our theoretical model implies the impact of market thickness on supply ratios between regions dwindles as the market thickens. Second, our data show that as Lyft grows, its relative outflows in different boroughs become less responsive to its size and settle at values that are close to those of Uber (in fact, these patterns are what identify the critical size). Taking a non-linear least squares approach, we estimate this minimum required size at 3.29M or 3.65M rides/month, depending on the functional form specification. This is close to Lyft's current size, substantially larger than Via's, and substantially smaller than Uber's. We suggest that such a minimum size should be had in mind in the current policy debate around downsizing rideshare platforms in NYC, given that falling short of it may distort the geographical distribution of supply at the expense of the outer boroughs. Our method can be used fairly straightforwardly and with limited data requirements to estimate such critical platform sizes for other metropolitan areas.

This paper has certain limitations. On the empirical front, although our reduced-form approach makes it more practical to implement, it also comes at a cost: unlike complementary studies, (Brancaccio et al. (2019a); Buchholz (2018); Frechette et al. (2019), etc.), our framework does not deliver welfare analysis. On the theory side, our model is static and abstracts away from dynamic (surge) pricing. It also abstracts away from platform competition. Although we provide empirical evidence that suggests the factors from which our theory abstracts are not likely to be first order in the understanding of geographical supply inequities, future studies can extend the model in those directions.

2 Related Literature

Our paper relates to multiple strands of the literature: (i) the recent and growing literature on the empirical analysis of geographical distribution of supply, and its possible distortion from that of demand, in spatial markets; (ii) the literature that studies the effects of market thickness in two-sided markets; and (iii) the literature that analyzes various aspects of the ridesharing market.

The empirical literature on the spatial match between supply and demand is new and small. To our knowledge, Buchholz (2018); Brancaccio et al. (2019c) are the only papers directly examining this issue, and papers such as Frechette et al. (2019); Brancaccio et al. (2019a,b) look at related problems. They extend the empirical techniques in the matching literature (see Petrongolo and Pissarides (2001) for a survey) in order to structurally infer the size of unobserved demand (e.g., passengers searching for rides) in different locations of a decentralized-matching market, when only the size of supply (e.g., available drivers) and the number of demand-supply matches (e.g., realized rides) are observed. Our paper is complementary by offering a reduced form approach to study the geographical distribution of unobserved potential demand, using data on the number of matches (rides) only. We will achieve this by noting that, in order to infer the magnitude of unfulfilled demand in a region, in addition to data on matches (rides) started in that region, one could leverage data on rides that started elsewhere but *ended* in the said region. Our method applies not only when the matching system is decentralized, but also when it is centralized. Also, it can detect relative under-supply in an area even if its passengers have learned, over the long run, not to search for rides.

Another subset of the literature on spatial markets that this paper builds on is the study of location decisions, resulting in agglomeration. Papers such as Ellison and Glaeser (1997); Ahlfeldt et al. (2015); Datta and Sudhir (2011); Holmes (2011); Miyauchi (2018) examine agglomeration of firms or residents. We add to this literature by arguing, empirically and theoretically, that agglomeration is also present in transportation markets. In addition, our comparative static theory results, which characterize how the extent of agglomeration is impacted by different factors, may be applied beyond transportation systems.

The second set of papers that we relate to is a large, mostly theoretical, literature on the impact of market thickness on the functioning of two-sided platforms in general (such as Akbarpour et al. (2017); Ashlagi et al. (2019)) and transportation markets in particular (such as Frechette et al. (2019); Nikzad (2018)). This literature, to our knowledge, has not examined how the spatial distribution of supply –and its (mis)alignment with that of potential demand– respond to a change in market thickness. Our paper focuses on this, both empirically and theoretically.

The third strand of the literature that our paper relates to is the set of papers on the functioning of transportation (in particular rideshare) markets. This strand itself can be roughly divided into (at least) two categories. One category is the group of papers focusing on this market as it relates to

labor economics. Chen et al. (2017) examine how much workers benefit from the schedule flexibility offered by ridesharing. Cramer and Krueger (2016) study the extent to which ridesharing, compared to the traditional taxicab system, reduces the portion of time drivers are working but not driving a passenger. Chen and Sheldon (2016) examine the reaction of labor supply to the introduction of ridesharing. Buchholz et al. (2018) estimate an optimal stopping point model to study the labor supply in the taxi-cab industry.

The second stream of papers on transportation/rideshare markets, to which our paper belongs, are those focusing on evaluating the performance of these markets as well as on market design aspects. Some of those papers, although related to our work in many ways, focus on questions that are inherently not spatial (examples are Cohen et al. (2016); Nikzad (2018); Lian and van Ryzin (2019); Cachon et al. (2017); Guda and Subramanian (2019)). Others study questions that are related to the spatial nature of the market (such as Castillo et al. (2017); Frechette et al. (2019)) but they do not examine the spatial distribution of supply and potential mismatches with demand. Many of the papers that do study geographical supply-demand (im)balance in transportation (such as Banerjee et al. (2018); Afèche et al. (2018); Castro et al. (2018)) focus on the short-run, intra-day, aspects. Some other papers (such as Buchholz (2018); Lagos (2000, 2003); Bimpikis et al. (2016); Shapiro (2018); Lam and Liu (2017),) however, examine long-term persistent mismatches. Our paper belongs to the latter group, and adds to it by studying the impact of market thickness.

Finally, it is worth noting that most of this literature has focused on the ways in which ride-share platforms improve upon the traditional taxi system, in particular due to their flexible pricing and superior matching algorithms (Cramer and Krueger (2016); Buchholz (2018); Frechette et al. (2019); Cohen et al. (2016); Shapiro (2018); Castillo et al. (2017); Castro et al. (2018); Lam and Liu (2017) among others). We add to this literature by comparing ride-share platforms to one another. We ask why is it that some rideshare platforms outperform others on some key issues, such as geographical reach, even though they all have superior technology relative to more traditional transportation systems? We conclude that a matching algorithm is not sufficient, and that other factors (i.e., adequate platform size) may be needed to ensure geographical reach. This enables our paper to quantitatively comment on the current policy debate regarding the appropriate sizes of rideshare platforms in NYC and other markets.

3 Data and Summary Statistics

We leverage two sources of data in this paper. The first data source is trip-level data that is publicly available from the Taxi and Limousine Commission (TLC henceforth) of New York City. For our main analysis, we use data on Lyft, Uber, and Via trips for July 2017 - December 2018. For each trip, we know the date and time of pickup and dropoff and a neighborhood indicator (again both for pickup and dropoff) partitioning NYC proper into 256 parts. Table 1 provides a summary of

this dataset.

Table 1: Summary of Ride counts across Platforms and Boroughs
(July 2017 - December 2018)

Platform	Bronx	Brooklyn	Manhattan	Queens	Staten Island	Toal
Pickups in 1000s of rides						
Lyft	3,518.66	19,336.19	23,956.76	10,925.45	437.72	58,174.79
Uber	23,567.13	57,993.56	96,084.76	34,417.17	1,853.50	213,916.12
Via	38.49	1,123.06	15,716.34	248.54	0.77	17,127.20
Dropoffs in 1000s of rides						
Lyft	3,601.44	19,364.25	22,954.42	11,796.50	458.18	58,174.79
Uber	24,202.12	58,182.93	91,752.35	37,897.13	1,881.59	213,916.12
Via	53.59	1,219.89	15,488.35	361.44	3.93	17,127.20

Also, Fig. 1 shows how platform size (measured in Millions of rides given per month) compares across platforms. Uber is by far the largest and Via by far the smallest. In terms of growth, Lyft has the highest percent growth, followed by Uber and Via, respectively. In addition, Fig. 2 shows a map of NYC with the population density of each of its five boroughs. Manhattan is the most densely populated, with a density of about 67 thousand/sq mile. Staten Island is the least dense borough, with a density of about 8 thousand/sq mile. Our empirical analysis will argue that in smaller rideshare platforms (i.e., thinner markets), supply gets more skewed toward higher density boroughs.

A second source of data in this paper consists of estimated pickup times and surge multipliers for all products of Uber and Lyft in NYC from late May 2015 to mid-June 2016. In the dataset, we observe the pickup times (as estimated by the platforms) and surge multipliers every 30 minutes in 195 locations across NYC proper for all products of Uber and Lyft–although in this paper we focus only on UberX (we term this Uber) and its Lyft-equivalent (Lyft). Hence, the unit of observation in this dataset will be the combination of (i) date, (ii) time of day, (iii) location, and (iv) platform. We use this dataset to find the right assumptions for our theoretical model and to justify those assumptions.

4 Empirical Analysis

Our empirical analysis has two main goals. First we develop an approach called "relative outflows analysis" to detect geographical mismatch between supply and demand. Second, we show empirical evidence for the role of market thickness (i.e., platform size) on the extent of such mismatch.

Figure 1: Platform Sizes for Uber, Lyft, and Via from July 2017 to December 2018. Vertical axis is on log scale.

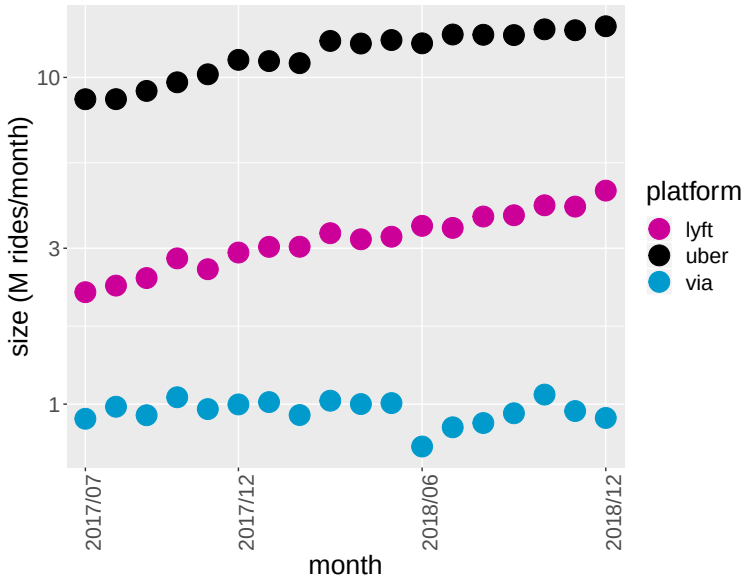
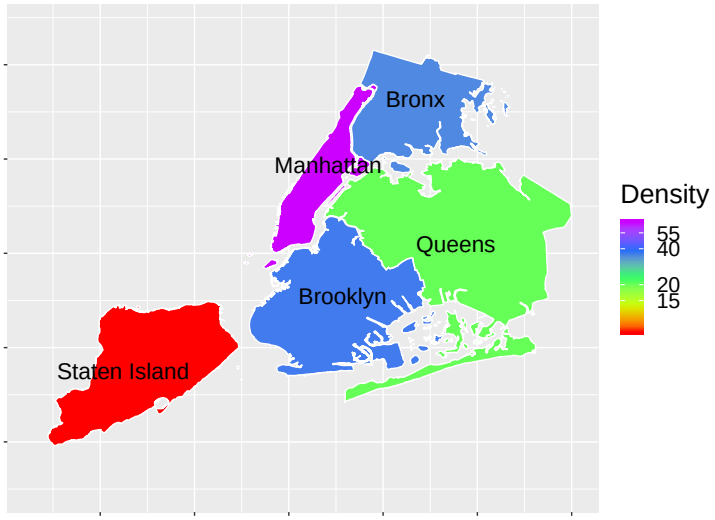


Figure 2: Population densities of the five boroughs of NYC as of April 2019 in thousands/sq mile. The color scale is logged.



4.1 Identifying Geographical Demand-Supply Mismatch

On a high level, our objective in this section is to identify the extent to which the geographical distribution of supply of ridesharing is "skewed away" from that of (potential) demand, leaving some regions under-supplied relative to others. To this end, we need to take two steps. First, we need to formally define what we mean when we say the geographical distribution of supply does not match that of demand (or, interchangably, when we say some areas are under-supplied relative to others; or there is geographical inequity in supply of rideshare services). Second, once equipped with an operational definition of demand-supply mismatch, we need to devise a strategy to empirically measure the extent of such mismatch. We now turn to these two tasks.

We say the geographical distribution of supply for rideshare platform k does not match that of its (potential) demand during time period d (e.g., a month or a day), if "access" to k's supply is heterogeneous across geographical regions i. By access to supply, denoted A_{ikd}, we mean the percent fulfillment of potential demand, measured in the following way:

$$A_{ikd} \equiv \frac{n_{ikd}}{\lambda_{ikd}} \tag{1}$$

where n_{ikd} is the total number of rides with platform k that originated in i during time period d; and λ_{ikd} is the "potential demand" for rides with k during d in i. We think of λ_{ikd} as having three components:

$$\lambda_{ikd} \equiv n_{ikd} + \lambda^{S}_{ikd} + \lambda^{NS}_{ikd} \tag{2}$$

That is, the total potential demand for rides with k in i during d is the sum of (i) n_{ikd}, the total number of customers who took rides with k from i during d, (ii) λ^{S}_{ikd}, the total number of those who searched for rides on their apps and decided to not take one, due to high prices or wait times, and (iii) λ^{NS}_{ikd}, the total number of those potential customers who decided not to search because they anticipated the rides would be too expensive and/or require very long wait times. In the literature, potential demand is usually assumed to consist only of the total number of those who search (e.g., see the definition of "those who seek rides" in Cohen et al. (2016)). That is, it excludes λ^{NS}_{ikd}. We insist on including λ^{NS}_{ikd} in spite of the empirical challenges its unobservability poses.[5] We believe excluding it could (substantially) understate the geographical inequity of supply, if customers in regions with high wait times/prices have responded by not going on their apps to search for rides.

Thinking of geographical demand-supply mismatch in terms of geographical inequity in A_{ikd} has some advantages. First, it is simple. It allows us to think of low access to rides in an area without having to specify whether the underlying reason of the low access is wait time, price, or both. This allows us to conduct our analysis without a requirement to (i) have data on prices and wait times or to (ii) specify a demand model that would capture how customers weigh wait times

[5] Brancaccio et al. (2019c) also take incorporate those who need transportation, but rationally decide against searching, into the pool of potential demand (or as they call it, "potential customers.")

and prices against each other. In spite of its simplicity, we find our definition relevant. Crucially, it helps deal with demand confounds in measuring supply-side differences across regions: if region i has fewer rides (or even rides/sq mile) than region j, it could either be due to lower demand in i or weaker supply in i. However, if our measure of access is lower in i than in j, we can interpret it as a supply-side difference given that A_{ikd} has the potential demand in its denominator. In addition, Section 5.5 will discuss two channels (one formally and one informally) through which geographical inequity in A across i is directly related to efficiency in the market.

With this definition in hand, we next turn to measurement. This is challenging because although we observe realized rides n_{ikd} in our data, the "unfulfilled" part of potential demand is unobserved. The amount of search for rides that did not lead to actual rides, λ^{S}_{ikd}, is usually only observed by rideshare platforms themselves. In addition, λ^{NS}_{ikd} is generally unobservable. In order to infer λ^{S}_{ikd}, the literature combines a model of matching with data not only on rides originating at i, but also on (observed or inferred) number of vacant cars; and then "inverts" the matching function to infer how many passengers must have searched for rides with k in i during d. In order to infer λ^{NS}_{ikd}, this matching model itself would need to be nested within a model of passenger-decision-making on whether to search.

To measure whether different areas have unequal access to rides, we propose a complementary method (to the matching function approach). Our approach requires data only on rides, and not on other variables such as vacant cars, wait times, or prices. Also, it does not require a model of matching or one of passenger decision making on whether to search for a ride. These features make our method readily implementable for not only academics and rideshare companies, but also policy entities with more limited available data and methodological sophistication. In addition, not requiring an explicit model of matching makes our approach applicable to all passenger-transportation markets irrespecitve of whether the matching mechanism is centralized (as in rideshare) or decentralized (as in taxicabs). Below, we implement the method.

The "Relative Outflows" Method. Our identification strategy, *relative outflows analysis*, leverages the information that inter-borough rides can reveal about unequal access to supply across regions. The basic idea is that if passengers use platform k to exit area i persistently and substantially less often than they do to enter it, then k must be under-supplied in i relative to outside of i. This conclusion is stronger if for other platforms k', such gap between inflows and outflows does not exist.

We start by making an approximation. Observe that demand for trips from region i can be for inter-region rides (denoted $\lambda^{\rightarrow}_{ikd}$), which *exit* i, or within-region rides (denoted $\lambda^{\circlearrowleft}_{ikd}$), which remain in i. Thus, overall demand for trips from region i can be written as: $\lambda_{ikd} \equiv \lambda^{\rightarrow}_{ikd} + \lambda^{\circlearrowleft}_{ikd}$ and

similarly for n_{ikd}. We then approximate the access as $A_{ikd} \equiv \frac{n^{\rightarrow}_{ikd} + n^{\circlearrowleft}_{ikd}}{\lambda^{\rightarrow}_{ikd} + \lambda^{\circlearrowleft}_{ikd}} \approx A^{\rightarrow}_{ikd} \equiv \frac{n^{\rightarrow}_{ikd}}{\lambda^{\rightarrow}_{ikd}}$.[6] Thus, our objective will be inferring geographical heteroteneity in "access to inter-borough rides" $A^{\rightarrow}_{ikd}$. Our inference will rely on two assumptions:

Assumption 1. *The frequency with which passengers migrate within the city (i.e., change where they live) is negligible relative to the frequently with which they take rides.*

Our second assumption has to do with the value of outside options. We first present a stronger version of it in Assumption 2, and later weaken it in Assumption 3.

Assumption 2. *The quality of outside transportation options is geographically homogeneous.*

Assumptions (1) and (2) imply that for any platform k and area i, the potential demand for rides exiting i is the same as the potential demand for rides entering it.[7] That is:

$$\lambda^{\rightarrow}_{ikd} = \lambda^{\rightarrow}_{i^c kd} \tag{3}$$

where i^c is the complement of region i, with respect to the whole city. It follows that:

$$\frac{A^{\rightarrow}_{ikd}}{A^{\rightarrow}_{ik^c d}} = \frac{n^{\rightarrow}_{ikd}}{n^{\rightarrow}_{ik^c d}} \equiv RO_{ikd} \tag{4}$$

That is, access to inter-region rides in i relative to that in the rest of the city can be directly measured by the observed *relative outflow* of rides at ikd, denoted RO_{ikd}. Importantly, we do not need to observe the potential demands. If $RO_{ikd} < 1$, it means the same population who choose k on their way into region i are on average systematically less likely to choose it over other options on their way out (Given that these are the same population, the difference in the flows cannot be attributed to differences in preferences for brands/modes of transportation, etc.). As an example, Lyft's relative outflow was 0.64 in Staten Island during July 2017. Using assumptions (1) and (2), we interpret this as Lyft being under-supplied in Staten Island compared to the rest of NYC proper because access to (outgoing) rides in that borough is 0.64 of the rest of the city. Note that this is the long-run under-supply because the number 0.64 combines the effect of those passengers who, on their way out of Staten Island, go on the Lyft app and do not find rides and those passengers who do not go on the app in anticipation of possible failure to find a ride.

Assumption 2 can be weakened to reflect the possibility that outside transportation options may be heterogeneous across regions:

[6] Later, we will discuss the consequences of this approximation.

[7] Of course for this to hold, the length of time period d should be long enough, (e.g., at least a day) so that for every trip there is a "trip back". This is why our approach applies to long-term rather than intra-day geographical imbalances between supply and demand.

Assumption 3. *The impact of outside transportation options on (potential) demand may be region-specific but cannot be platform-region specific.*

That is, if we replace Assumption 3 for Assumption 2, we will not get $\lambda^{\rightarrow}_{ikd} = \lambda^{\rightarrow}_{i^c kd}$ anymore. However, for any two platforms k and k', we get:

$$\frac{\lambda^{\rightarrow}_{ikd}}{\lambda^{\rightarrow}_{i^c kd}} = \frac{\lambda^{\rightarrow}_{ik'd}}{\lambda^{\rightarrow}_{i^c k'd}}$$

Thus, although eq. (4) will not hold anymore, we can obtain the following:

$$\frac{\frac{A^{\rightarrow}_{ikd}}{A^{\rightarrow}_{ik^c d}}}{\frac{A^{\rightarrow}_{ik'd}}{A^{\rightarrow}_{ik'^c d}}} = \frac{RO_{ikd}}{RO_{ik'd}} \tag{5}$$

In other words, according to assumptions (1) and (3), if Lyft's relative outflow of 0.64 in Staten Island is a mere reflection of attractive outside transportation options (rather than lower access to Lyft's supply on the island), then such outside options should also bring Uber's relative outflow down to 0.64 in that region in the same time period. This assumption, therefore, combined with multi-platform data, will allow us to still infer geographical supply inequities even if outside options are not geographically uniform. In the case of this example, Uber's relative outflow during July 2017 was 0.95, suggesting, according to eq. (5), that access to supply for Lyft in Staten Island must have been low (relative to Uber).[8]

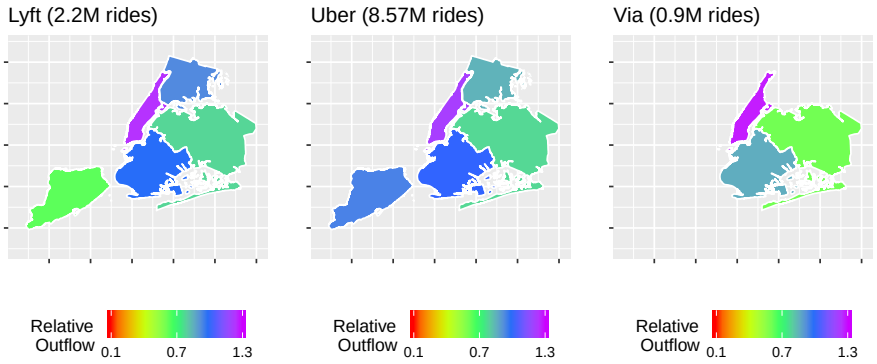
Using Relative Outflows to Quantify Geographical Inequity. Having described the assumptions behind the relative outflows analysis and its interpretation, we now turn to the execution of the method and interpretation of the results. Figure 3 depicts relative-outflows for each platform during July 2017 for all boroughs in which the platforms were operating at that time. The figure indeed suggests some areas are under-supplied relative to others. To formally analyze this, we run a regression detailed in Eq. (6). The results from this regression are reported in Table 2.

$$\log(RO_{ikd}) = \text{FE}_i + \gamma_i^{Lyft} \times \mathbf{1}_{k=\text{Lyft}} + \gamma_i^{Via} \times \mathbf{1}_{k=\text{Via}} + \varepsilon_{ikd} \tag{6}$$

Like Fig. 3, Table 2 shows geographical inequity in supply of rideshare across NYC boroughs. For instance, under assumptions (1) and (2), an arriving potential demand for an (interborough)

[8] It is worthwhile to understand what Assumption 3 is ruling out. As an illustration of this point, we are ruling out the following possibility: Lyft users who travel in and out of Staten Island tend, substantially more than their Uber-user counterparts, to be from a demographic group that travels into Staten Island during hours when ferries are not available but travels out during hours when ferries are working. If such one-directional differences exist among platforms, it can undermine our interpretation that the difference between relative outflows of Uber and Lyft in Staten Island (i.e., 0.64 vs. 0.95) comes from relative under-supply of Lyft. One way to test this is to look at hours of the day and check whether relative outflow of Lyft is below that of Uber persistently within the day or only at certain times of day. We carry out such a test and detail it in Appendix B.

Figure 3: Relative Outflows for Uber, Lyft, and Via across NYC boroughs in July 2017



Uber ride is $e^{0.192} = 1.21$ times more likely to be served in Manhattan than it is elsewhere, and this difference is statistically significant (i.e., persistent over the month).[9] Under assumptions (1) and (3), however, the fixed effects coefficients cannot be interpreted on their own. Nevertheless, the interaction coefficients may be interpreted in a cross-platform manner. For instance, relative to Uber, Lyft is under-supplied in Staten Island by a factor of $e^{0.388} = 1.47$.

Our method has some caveats. First, it applies only to passenger transportation markets. Given the important role of Assumption 1, a transportation market for internationally traded goods (such as the market studied by Brancaccio et al. (2019a,b)) cannot be studied using our method. Second, our method does not recover the absolute value of percent fulfillment of demand in each region. Rather, it shows how each region compares against the rest of the city. Third, although our approach prevents understating geographical inequity in supply due to passengers not searching, it is still prone to understating the extent of inequity, for two reasons: (i) if a passenger decides to forgo a trip to, say, Staten Island because the ride back will be hard to find, it will not show in relative outflows; and more importantly, (ii) passengers are perhaps more likely to forgo a within-borough ride if the wait time is high than they are to forgo a longer, inter-borough, one.[10]

To sum up the analysis thus far, we developed the relative outflows methods and applied it to rideshare in NYC to demonstrate that geographical inequity in supply exists among the NYC boroughs. Our next section studies the possible mechanisms leading to this inequity.

[9] Although we find Assumption 2 a strong one, we do believe it is useful to mention how the results would be interpreted under assumptions 1 and 2. The reason is that there are several anecdotes suggesting that outside transportation options are in fact more accessible in Manhattan. Therefore, if anything, the results may understate the under-supply of Uber in the outer boroughs relative to Manhattan.

[10] In other words, approximating A_{ikd} by $A^{\rightarrow}_{ikd}$ may bias our measure towards 1, thereby understating the heterogeneity.

Table 2: Relative Outflows Regression, July 2017

Dependent variable: log(Relative Outflow)

	Estimate	(SE)
Bronx	-0.164^{***}	(0.020)
Brooklyn	0.015	(0.020)
Manhattan	0.192^{***}	(0.020)
Queens	-0.256^{***}	(0.020)
Staten Island	-0.054^{***}	(0.020)
Lyft × Bronx	0.092^{***}	(0.028)
Lyft × Brooklyn	-0.032	(0.028)
Lyft × Manhattan	0.029	(0.028)
Lyft × Queens	0.002	(0.028)
Lyft × Staten Island	-0.388^{***}	(0.028)
Via × Brooklyn	-0.169^{***}	(0.028)
Via × Manhattan	0.052^{*}	(0.028)
Via × Queens	-0.239^{***}	(0.028)
Observations	403	
R^2	0.832	
Adjusted R^2	0.826	
Residual Std. Error	0.112 (df = 390)	
F Statistic	148.536^{***} (df = 13; 390)	

*p<0.1; **p<0.05; ***p<0.01

Note: Uber is the omitted group. These results suggest that there is geographical inequity in supply of rideshare across boroughs of NYC.

4.2 What Leads to Spatial Mismatch between Demand and Supply?

In this section, we ask what can cause geographical supply inequity. We start in Section 4.2.1 by providing empirical evidence for the role of market thickness (i.e., platform size) in impacting the geographical inequity in supply of ridesharing in NYC. Next, in Sections 4.2.2 through 4.2.4, we (i) dig deeper into whether the inequity in access is caused by prices or wait times and (ii) offer suggestive evidence in support of a driver-behavior based mechanism and against some alternative mechanisms. All of our results in this section are crucial in motivating our theory model.

4.2.1 Evidence for the Role of Market Thickness (i.e., Platform Size)

The results in Fig. 3 and Table 2 not only show demand-supply mismatch exists, but also demonstrate a pattern. Smaller rideshare platforms seem to be under-supplied in boroughs with lower population densities. This hypothesis is strengthened once we compare those relative outflows (which were from July 2017) with those from a year later, June 2018. Fig. 4 does this job and shows multiple interesting patterns, including: (i) Lyft's relative outflow distribution becomes more similar to Uber's as Lyft grows in size, and (ii) once active in Staten Island, Via has a very small relative outflow of 0.13 there.

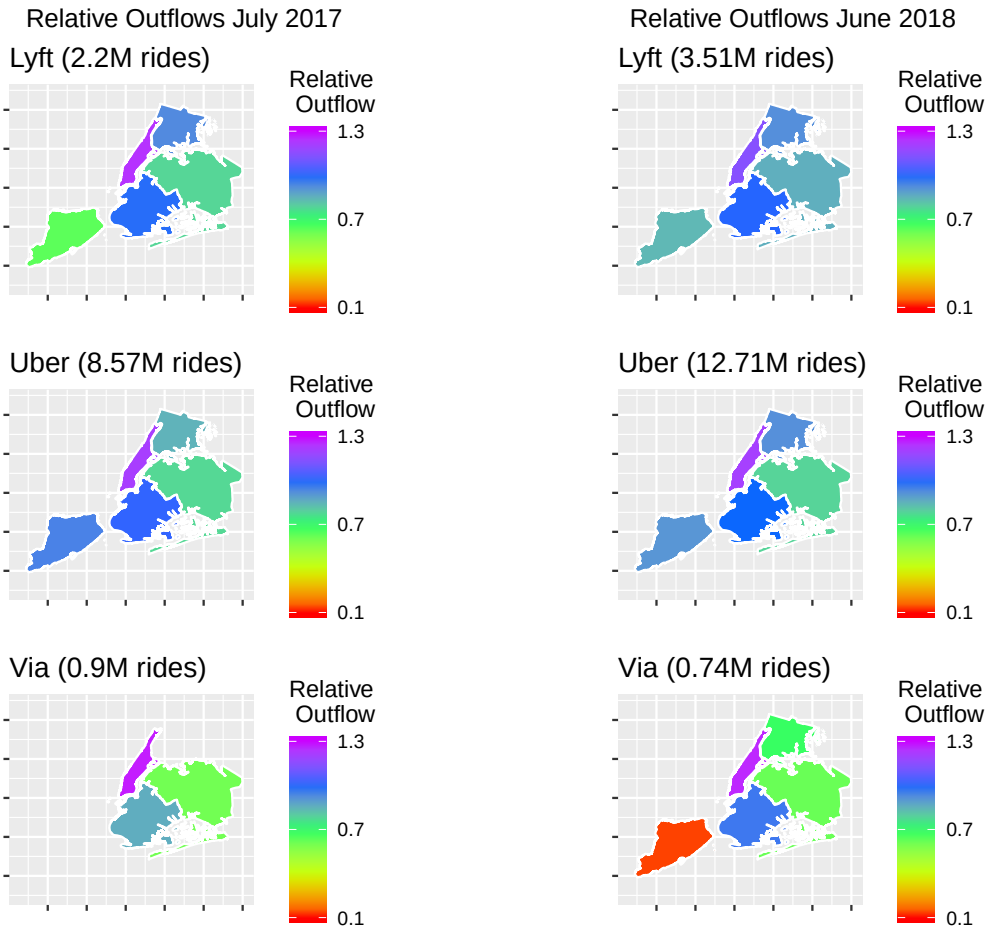
To formally examine the association between market thickness (platform size) and low access to supply in lower density boroughs, we run the following regression, and we run it on monthly relative-outflow data from all three platforms from July 2017 to December 2018.

$$RO_{ikd} = \alpha_0 + \alpha_1 \log(\rho_i) + \alpha_2 \log(S_{kd}) + \alpha_3 \log(S_{kd}) \log(\rho_i) + \nu_{ikd} \tag{7}$$

where RO_{ikd} is the relative outflow for platform k at borough i on date d. Also ρ_i is the population density of borough i as of April 2019 as a proxy for demand density.[11] Finally, S_{kd} is the size of platform k on date d, which is measured by the total number of rides given by that platform in NYC during the month in which date d occurs. Tables (3) and (4) report the results from this regression. The first table reports results when we either do not include any fixed effects in the regression or we do have fixed effects but they are not interacted (that is, platform fixed effects, year-month fixed effects, or borough fixed effects). Table 4, however, incorporates a much richer set of fixed effects. It starts, in its first three columns, fixed effects on (i) platform interacted by year-month, (ii) borough interacted by platform, and (iii) borough interacted by year-month. It then incorporates these three pairs into one single regression. Finally, the last column has interaction

[11] We use population densities to proxy for unobservable densities of (potential) demand in boroughs. A more though approach, perhaps, would be to use pickup densities instead and construct an instrument for it using population densities. We believe, however, that such additional complication would not substantially add to our understanding of the market. This is in part because population densities seem to be very good proxies, given that the rank-order of population densities across boroughs closely matches that of rides.

Figure 4: Relative outflows for Lyft, Uber, and Via†



(a) (b)

†: Panel (a) is July 2017 and Panel (b) is June 2018

fixed effects among boroughs, platforms, and years (not year-month in this column).

Table 3: Relative Outflow Regression with Single Fixed Effects

	Dependent variable: Relative Outflow			
	(1)	(2)	(3)	(4)
Constant	-7.371^{***}	-	-	-
	(0.142)			
log(population density)	2.154^{***}	2.222^{***}	2.141^{***}	-
	(0.041)	(0.040)	(0.041)	
log(size)	0.492^{***}	0.483^{***}	0.490^{***}	0.448^{***}
	(0.009)	(0.014)	(0.009)	(0.008)
log(population density)×				
log(size)	-0.126^{***}	-0.130^{***}	-0.125^{***}	-0.113^{***}
	(0.003)	(0.003)	(0.003)	(0.002)
Fixed Effects†	None	P	YM	B
Observations	7,709	7,709	7,709	7,709
R^2	0.595	0.624	0.598	0.725
Note:				*p<0.1; **p<0.05; ***p<0.01

†: P:Platform, B:Borough, YM:Year-Month

There are two coefficients of interest for us in these tables. First, the coefficient on population density is positive and significant in all specifications. It implies that a higher relative outflow should be expected for boroughs with higher population densities.[12] Second, and more important, is the interaction coefficient which is negative and also statistically significant across all the specifications in both tables. This coefficient indicates that as the platform gets larger, the disparity between relative outflows of low and high density boroughs shrinks. These results are robust to the inclusion of such a rich set of fixed effects because the association between small platform size and under-supply in low density areas is reflected in multiple sources of variation.[13] For instance (i) looking at variation within platforms over time, one can see Lyft's relative outflows in July 2017 are less uniformly distributed compared to that in June 2018; or (ii) by looking at variation across platforms

[12]Obviously, this coefficient cannot be separately identified from borough fixed effects or interactions of borough fixed effects with other effects.

[13]On top of this rich set of fixed effects specifications, we also study different functional form assumptions. In equation eq. (7), we log population density and platform size but not the relative outflow. One could think of eight different specifications here depending on which subset of these three variables are logged. We took all of these 72 regressions (8 functional form assumptions × 9 fixed effects specifications) and the interaction coefficient is always negative with the corresponding p-value never exceeding 4×10^{-7}.

Table 4: Relative Outflow regression with Interaction Fixed Effects

	Dependent variable: Relative Outflow				
	(1)	(2)	(3)	(4)	(5)
log(population density)	2.182***	-	-	-	-
	(0.040)				
log(size)	-	0.596***	0.444***	-	0.402***
		(0.036)	(0.008)		(0.061)
log(population density)×					
log(size)	−0.128***	−0.167***	−0.112***	−0.393***	−0.110***
	(0.003)	(0.011)	(0.002)	(0.023)	(0.018)
Fixed Effects†	P ×YM	B ×P	B × YM	(P × YM)	P × B × Y
				+(P × B)+(B × YM)	
Observations	7,709	7,709	7,709	7,709	7,709
R^2	0.629	0.829	0.738	0.841	0.835
Note:					*p<0.1; **p<0.05; ***p<0.01

†: P:Platform, B:Borough, YM:Year-Month, Y:Year

at a certain time, one can see Lyft's relative outflows during June 2017 are less uniformly distributed than Uber's during the same time period. According to these results, for instance, if a rideshare platform's size is cut in half, it would lead to about 10 percentage points higher loss of rides needed in Queens than it would in Manhattan.[14]

Although we lack a quasi-experimental variation, we interpret our results on the negative interaction coefficient causally. We do so because the robustness of the results allows us to rule out a wide range of alternative explanations for why there is an association between low platform size on the one hand and under-supply in less dense boroughs on the other. First, note that demand side confounds are ruled out by our choice of dependent variable. That is, if the dependent variable were the number of pickups by k in i, then, for instance, exceptionally few Via pickups in Staten Island (compared to other platforms) could be either because of low local demand for Via (e.g., Staten Islanders being unaware of Via, or more loyal to other brands) or because of limited access to supply. However, because our dependent variable is the relative outflow, under assumptions (1) and (3), such demand side differences across platforms cannot explain the negative estimates for our coefficient of interest α_3 in regression (7).[15]

[14]Of course the effect will be non-linear and its magnitude will itself depend on the size. For more details on this, see Section 6.

[15]In fact, the robustness of the results from Tables (3) and (4) can help rule out demand-side explanations for the negative α_3 estimates, *even when Assumption 3 is weakened.* The only way a negative and significant α_3 can

Our results in Tables (3) and (4) also help rule out some supply-side explanations for the negative α_3 estimates which can be considered alternative explanations to our market-thickness hypothesis. One such alternative is: "drivers of Lyft are less likely to live in Staten Island; hence, they are less available for pickups there." But this can only be consistent with our results from Figure 4 and Tables (3) and (4) if this difference between where Uber and Lyft drivers lived was time-varying and dwindled in 2018. Also, it must be that Via drivers are even more likely than both Uber and Lyft drivers to live in busier areas. Another alternative explanation for our results on α_3 would be "Uber and Lyft have different incentive mechanisms for their drivers in terms of where they are encouraged to drive." As we will discuss shortly, prices are not likely to have been of first order importance in differentially encouraging drivers of smaller platforms to drive in denser areas. Aside from this, for any price or non-price incentive mechanism to comprise an alternative explanation to the impact of market thickness, the following need to be true. It must be that the difference between Uber's and Lyft's incentive policies dwindled over time, in a manner correlated with the movement of their size ratio, but *not because* of the movement in their size ratio. Also, it must be that the correlation between the differential incentives provided by platforms and their size ratios is finer than yearly movements. This can be seen in column 5 of Table (4), where our coefficient of interest remains negative and significant even when we interact the platform-borough fixed effects by year dummies.

Although our analysis suggests a thinner market, all else being equal, should increase the relative under-supply in the outer boroughs, the mechanism through which this effect works is not clear. Our theory model in Section 5 will describe what we believe is the most parsimonious mechanism that can explain this phenomenon. Before turning to that theory, however, we present suggestive evidence that (i) prices do not play a first order role in making supply in smaller platforms (i.e., thinner markets) more concentrated in busy areas and (ii) platforms' possible competitive responses to each other do not seem to play a first order role either. We also present anecdotal evidence for the role of driver behavior in shaping the under-supply of rides in less busy areas. These pieces of evidence will have implications for our modeling decisions in Section 5.

4.2.2 What is the Impact of Prices?

In this section, we provide several empirical arguments to show that the under-supply of rideshare in less dense areas is not primarily caused by price differences across platforms and regions.

be interpreted as a demand side phenomenon is if there is a platform direction-specific difference among passengers. That is, for instance, if (i) Lyft passengers are more likely than Uber passengers to prefer public transportation; and (ii) in Staten Island, unlike other boroughs, public transportation is not more available. This is a relaxation of Assumption 3. But even this cannot explain the negative α_3 in column 2 from Table 4, which has platform-borough interaction fixed effects. The only way to interpret the negative α_3 in that column as a demand-side phenomenon would be to assume that such platform direction-specific differences are *also time varying* and diminish in 2018.

The first and shortest argument–but by no means the least powerful one–is that Via does not use surge pricing, unlike Uber and Lyft; however, as demonstrated before, Via has the most skewed relative outflows distribution among all three platforms. As previously shown, in 2018 when Via started operating in Staten Island, the relative outflows for Via in Staten Island were much smaller than both other platforms (it was, for instance, about 0.13 in June 2018). This suggests that price is not likely to be the main driver for the documented geographical inequity.

We also use our data on surge multipliers and estimated pickup times (ETAs) from Uber and Lyft apps from late May 2016 to mid-June 2016 to offer suggestive evidence against the importance of prices. The time window of this data does not overlap with our rides data, unfortunately. However, we find the insights from the analysis of this data source useful for our modeling decision in the theory part of the paper. We conduct two analyses. One uses data on the entirety of NYC proper. The other focuses on Staten Island only, where the prices charged by Uber and Lyft are constant across time and locations.

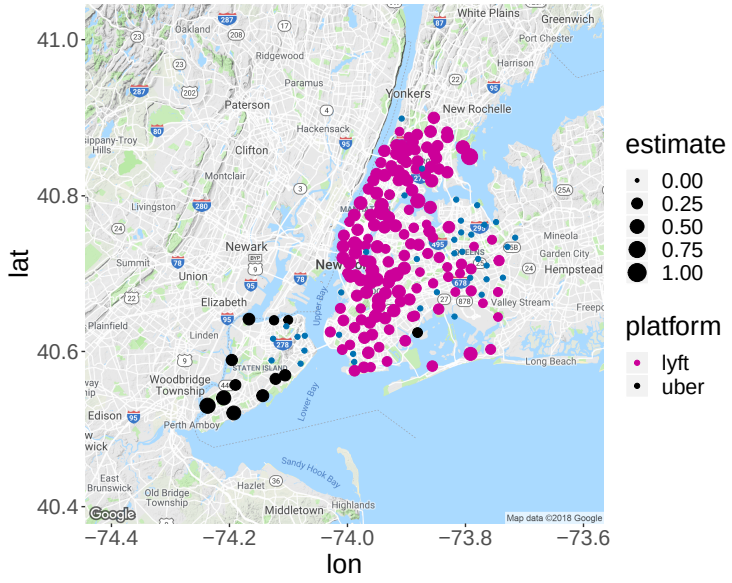
Analysis of NYC proper. This analysis accomplishes two tasks. First, we show that the geographical patterns in disparities between Uber's and Lyft's pickup times are similar to the patterns previously documented for relative outflows, whereas the same is not true of prices. This suggests that potential demand for Lyft rides in less dense areas tends to go unfulfilled more than Uber's does, mainly because Lyft's pickup time is high in those regions, rather than because Lyft's prices are high. Second, we offer evidence suggesting that the geographical disparities between Lyft's and Uber's pickup times are *not caused* by geographical disparities in their prices. These two analyses, together, suggest prices matter little, either directly or indirectly, in shaping the observed geographical supply inequity.

As for the first task, we start by documenting the geographical disparity between Uber's and Lyft's pickup times, using the following regression:

$$\log(t_{ikdt}^{\text{pickup}}) = \Delta_i + \delta_i \times \mathbf{1}_{k=\text{Lyft}} + \epsilon_{ikdt} \tag{8}$$

where t_{ikdt}^{pickup} is the pickup time for platform k in location i on date d and time t. Also, Δ_i and δ_i are, respectively, location fixed effects and the interaction between location and a "Lyft dummy." The coefficients of interest are δ_i. Wherever δ_i is positive and significant, it means Lyft, on the average, has a longer pickup time than Uber. The reverse is true wherever the estimated δ_i is negative and significant. Fig. 5 visually depicts the regression results for δ_i values. Pink points are locations in which Lyft is expected to arrive faster, whereas in black ones, Uber has a shorter pickup time. Blue points are those at which δ_i is not statistically significant. Due to the high number of coefficients (195 separate δ_i values,) there is a risk of spurious correlation. Thus, we focus on coefficients significant at the 99.9999% level, instead of the more common 95%.

Figure 5: Platform Pickup Time Regression Estimates $(\delta_i)^{\dagger}$



†: Lyft estimated to arrive faster in pink areas and Uber in black areas. In blue areas, neither platform is faster than the other in a statistically significant way at the confidence level of 99.9999%.

Interestingly, the comparison between Lyft's and Uber's pickup times varies geographically, in a similar way to the comparison between the relative outflows. In particular, Uber has the highest advantage in Staten Island. Outside of the island, Lyft tends to have a smaller pickup time, with both the significance and magnitude dwindling as we more to lower population density areas of the city.[16] This suggests that in lower density areas, passengers may forgo Lyft rides due to high pickup times. The same, however, is not true of surge multipliers. Table 5 shows the average surge multipliers for the two platforms across different regions. As shown in the table, although Lyft's surge factor is about 0.01 higher than Uber's in black (less densely populated) areas, it is about 0.05 higher in pink (more densely populated) ones. This suggests that it is unlikely that Lyft rides are forgone in lower density areas due to high price.

Our second task is to offer evidence that differences between the two platforms' prices do not cause Lyft's pickup times to be so much higher than Uber's in black areas of Fig. 5 (roughly Staten Island) and the reverse to be true in pink areas. To see this, note that Lyft's pickup time is 196.21 sec longer than Uber's in black areas, while it is 77.78 sec shorter in pink ones. This means that the total "difference in difference" between the two platforms across pink and black areas is

[16]The comparative geographical patterns of UberX and Lyft are not impacted by the possibility that one platform underestimates its pickup times relative to the other, as long as such underestimation happens in all locations. As a robustness check, we "de-averaged" all of the pickup times by dividing all UberX and Lyft pickup times by the average UberX pickup time and average Lyft pickup time, respectively. Doing the rest of the analyses in this section based on those de-averaged numbers did not change the results.

Table 5: Pickup Times (in seconds) and Surge Multipliers by Area†

	Black Area	Pink Area
Lyft Pickup Time	652.80	247.66
Uber Pickup Time	456.59	325.44
Lyft Surge	1.012	1.102
Uber Surge	1	1.052

† Area Black or Pink based on Figure 5

$196.21 - (-77.78) = 274.00$ seconds.

We now ask how much of this difference in difference is *caused* by geographical differences between the platforms' prices–that is, how much this difference in difference would change if the two platforms had a constant surge factor of 1 across locations and time. In order to estimate this object, one would need a causal estimate of surge multipliers on pickup times. Although we do not have such an estimate, we borrow it from Cohen et al. (2016), who use a regression discontinuity design to estimate demand in ridesharing. They estimate that in 2015 in NYC, Chicago, Los Angeles, and San Francisco, a 0.1 increase in the surge multiplier for UberX causes the pickup times to decrease by 7.7 seconds on average. If we assume their estimate is also valid for our context (Lyft and UberX in May and June of 2016, in NYC only,) then a constant surge multiplier of 1 across platforms, locations, and time would decrease the average difference in difference in pickup times by the following amount:

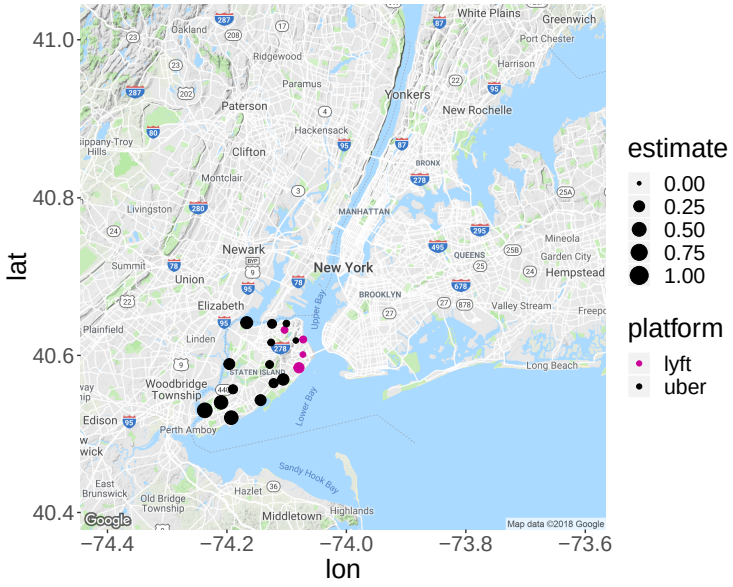
$$\frac{7.7sec}{0.1} \times [(1.102 - 1.052) - (1.012 - 1)] = 2.85sec$$

This is approximately only 1% of the observed average difference in difference in pickup times (which was 274 sec), suggesting that price differences between platforms are too little to explain the geographical disparities between the platforms' pickup times. Of course our context is not exactly the same as the context in Cohen et al. (2016). Nevertheless, given that with this coefficient, prices explain only 1% of the geographical disparity in pickup times, they would still explain fairly little even if the true value of the coefficient in our context is much larger than that in Cohen et al. (2016).

To sum up, our analysis of UberX and Lyft pickup times and surge multipliers suggests that prices neither play a direct substantial role, nor an indirect one, in the under-supply of Lyft in less dense areas.

Analysis of Staten Island from 2am - 6am. In this section, we focus our attention to Staten Island only (as opposed to all of NYC), and to the hours of 2am - 6am only (as opposed to 24 hours). Both platforms do minimal or no surge pricing at this location and time interval. Therefore, if there

Figure 6: Platform Pickup Time Regression with no Surge in Staten Island.



†: Lyft estimated to arrive faster in pink areas and Uber in black areas.

is a large disparity between Uber's and Lyft's estimated arrival times across different locations within Staten Island between 2am and 6am, it cannot be caused by price differences between the two platforms. This will provide the same empirical evidence that our analysis of NYC proper offers, except it does not rely on estimates from Cohen et al. (2016), which came from a slightly different context.

Figure 6 is thus restricted to 2am - 6am in Staten Island (we show all points, instead of just the statistically significant ones).

Table 6: Pickup Times (in seconds) and Surge Multipliers by Area‡

	Black Area	Pink Area
Lyft Pickup Time	682.62	481.39
Uber Pickup Time	523.04	535.90
Lyft Surge	1.003	1.003
Uber Surge	1.000	1.000

‡ Area Black or Pink based on Figure 6

Table 6 summarizes the surge-factor and arrival-time comparisons across platforms and areas. It shows that while prices are constant across locations, Lyft's arrival time rapidly increases as we move toward less densely populated parts of the borough, whereas Uber's is relatively stable. As illustrated in the figure, there is almost no geographical price disparity. Uber never does surge

pricing, and Lyft does it very infrequently (fewer than 10 instances in more than 2,000 observations) and does almost equally between pink and black areas (in fact, if anything, the average surge for Lyft across black areas is about 0.0001 higher than that in pink areas). However, the estimated arrival times are very different. Lyft's arrival time is about 54 seconds faster in pink areas, whereas Uber's is about 160 seconds faster in black areas. The difference-in-difference is about 214 seconds and is statistically significant, with a standard error of 18 seconds.

4.2.3 What is the Role of Platform Competition/Collusion?

Another possible reason for the relative under-supply of rides from smaller platforms in the outer boroughs could be strategic interactions among platforms. It is, in principle, conceivable that platforms collude by strategically divide the city amongst themselves geographically in order to avoid direct competition against one another. In practice, however, we show that this does not seem to be the case in the NYC rideshare market. We argue this by observing that such collusion would have implications that are not empirically supported once we examine the patterns from our data. First, if platforms strategically send their drivers to different parts of the city, it would be natural to expect that prices are among the main levers platforms use to carry out this mission. This is, however, not empirically supported for at least two reasons: (i) Via does not do surge pricing; (ii) by the logic in Section 4.2.2 that prices do not seem to play a major role in Lyft's relative under-supply in less busy areas.

Second, if platforms are strategically dividing the market geographically, it would be natural to expect data patterns indicating that each platform is focusing on a certain area. This is not empirically supported either. All platforms in our data have higher ride densities, higher relative outflows, and lower ETAs in denser parts of the city than elsewhere. It is only the *slope* of decline (for rides and relative outflows) or increase (for ETAs) that is steeper for smaller platforms.[17] We cannot think of a form of strategic division of the market that would lead to this pattern.

Finally, a strategic division of the city by platforms would have implications for which platforms would focus on which areas. It would be natural to expect larger and more powerful platforms, which enjoy a first-mover advantage, to take the more attractive regions, and for the newcomers to find niche markets. This is not supported by the observation that the supply of smaller platforms is more skewed towards Manhattan (and, in general, busier areas) relative to those of larger platforms. For instance, Via started its business from Manhattan; and even when it became active in a low density borough such as Staten Island, it had a relative outflow of only 0.13.

[17] See Table 1, Fig. 4, and Table 5.

4.2.4 The Role of Driver Behavior

There is anecdotal evidence that drivers behave in a manner closely in line with our empirical results in Tables (3) and (4). In Appendix A, we document evidence from online rideshare forums that (i) drivers tend to avoid less busy areas because they consider pickups too far away and (ii) the problem with distant pickups in less busy areas, and hence, the avoidance of those regions by drivers, is a more pronounced problem for Lyft than it is for Uber. In addition to this anecdotal evidence, we have run an analysis (available upon request) of data from a ride-share platform in Austin to obtain direct empirical evidence on the impact of pickup times on driver behavior. Controlling for prices and idle times and resolving a series of endogeneity issues, our analysis obtains evidence that drivers tend to avoid the outer areas of the city, and that pickup times play a first order role in their decisions.[18]

To sum up, we find the suggestive evidence provided in Sections 4.2.2 through 4.2.4 strong enough to motivate our decisions in the theoretical analysis to (i) abstract away from prices, and (ii) study a monopolist platform instead of competition, but instead (iii) be more general than the literature when it comes to the spacial aspects of the market, in order to properly capture the role of pickup times on drivers' location decisions. We turn to that theoretical analysis next, which builds upon the results in Section 4.2.1, and with assumptions guided by the results in Sections 4.2.2 through 4.2.4.

5 Theoretical Model

Our theoretical model complements the empirical analysis in at least two ways. First, it describes a mechanism through which a thinner market (i.e., smaller platform size) can, *ceteris paribus*, lead to under-supply of the rides in less dense areas. Our empirical analysis suggests such a mechanism should be there, but the theoretical literature on spatial markets is silent on the relationship between market thickness and the geographical inequity of supply. The second role of the theoretical model is to produce further results that could help enrich the empirical policy analysis. For instance, as we will show in this section, our theoretical model suggests that the impact of platform size on the geographical distribution of supply will satiate once the platform size becomes large enough. We will feed this insight back to the empirical analysis in Section 6 in order to estimate that minimum adequate size, which might be of interest to policymakers.

[18] In spite of being otherwise rich, this Austin data is only from a single platform. This puts it at a disadvantage, compared to our NYC data for our empirical analysis. As such, we chose NYC for the main analysis in the paper. It is worth noting that the analysis of the Austin data also shows that relative outflows are substantially higher in busier parts of the city.

5.1 Setup

We model a market with regions $i \in \{1, ..., I\}$ with a monopolist ridesharing platform serving them. The regions (which, depending on the application, would could think of as neighborhoods, boroughs, etc.) are modeled as circumferences of circles, a la Salop. Regions are assumed to have the same size.[19] In each region, passengers arrive at a rate λ_i per unit of time. Without loss of generality, we assume $\forall i < j : \lambda_i \geq \lambda_j$. Also, λ represents the vector $(\lambda_1, ..., \lambda_I)$. Each arriving passenger's location is uniformly distributed on the circumference of the circle. There are a total N drivers who work for the platform.

Our model is a one-shot game among drivers in which they simultaneously and independently choose which of the regions to drive in. Once they choose their regions and n_i drivers pick region i, we assume for simplicity that they are uniformly distributed across the region (i.e., the circumference of circle i).[20] Drivers are matched to arriving passengers via a centralized matching system. Each driver's "range" or "catchment area" will be the arc consisting of all the points on the circle that are closer to that driver than they are to any other driver in region i. Each driver picks up the first passenger that arrives within that driver's catchment area. In practice, ridesharing platforms implement a similar matching rule (Frechette et al. (2019) use a similar approach to model a centralized matching market). The game finishes once all drivers have picked up their passengers.[21]

Each driver chooses a region to drive in, minimizing his or her expected total wait time, which has two components. First, the driver in a specific location must wait for demand realization, i.e., the arrival of a customer in the catchment area. We term this the **idle time**. Second, the driver needs to travel to the exact location of the customer to pick her up, and we call this the **pickup time**. The wait time is thus comprised of these two different components, which have divergent impacts on the equilibria in ridesharing markets.

The circular model of regions is illustrated in Figure 7. Suppose the disutility to a driver from traveling a full circumference to pick up a passenger is t' times that of one minute of idle time.[22] The platform allocates an arriving customer to the closest driver. Because drivers are situated at equidistant points on the circumference of the region, their catchment areas include half the distance to their nearest neighbors on both sides. The idle time expected for a customer to arrive

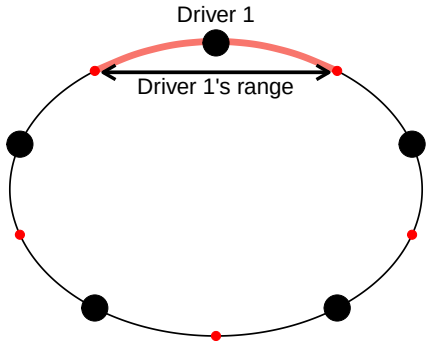
[19] The equal size assumption is not limiting since one can, in principle, think of a larger region as equivalent of multiple of our uniform-sized regions i.

[20] While we do not model the locational choice of drivers within the region, it is fairly easy to see that equidistant positioning location from neighbors is an equilibrium. While there might be other locational choices that might also be equilibria, we focus on the equidistant positioning equilibrium.

[21] Note that by not tracking the destinations, this model does not capture relative outflows. We do not see this as a weakness, given that the role of relative outflows in the empirical section was to help with the identification of λ_i values, which we assume known in the theory model.

[22] Thus, t' can be thought of as aggregating (i) how long it takes to travel the circumference, and (ii) how much more than idle time drivers dislike pickup time, due, perhaps, to feul costs.

Figure 7: Illustration of Circular Model of each Region i and Driver Allocation



in the driver's area is $\frac{n_i}{\lambda_i}$. The distance between drivers is $\frac{l}{n_i}$ where l is the circumference. Since consumer location is uniform, the distance a consumer will be from the driver along the arc is distributed $d \sim U[0, \frac{l}{2n_i}]$, implying that the expected distance is $E[d] = \frac{l}{4n_i}$. Thus, the (cost of) expected pickup time is $\frac{t'}{4n_i}$.

We have the (cost of) expected total wait time $W_i(n_i)$ defined from the driver's perspective as:

$$\underbrace{W_i(n_i)}_{\text{Total Wait Time}} = \underbrace{\frac{n_i}{\lambda_i}}_{\text{Idle Time}} + \underbrace{\frac{t'}{4n_i}}_{\text{Pickup Time}} = \left(\frac{n_i}{\lambda_i} + \frac{t}{n_i}\right) \tag{9}$$

where $t = \frac{t'}{4}$. Observe that idle time increases in the number of drivers n_i since a given level of customer demand is allocated across all the drivers present in the region. On the other hand, the ridesharing platform allocates each customer to the closest driver. Thus, the pickup time decreases in n, since with a greater number of drivers, each driver is more likely to be allocated a passenger closer to him. In other words, the presence of each driver in region i has a negative externality on other drivers in i by increasing their expected idle times and a positive externality on them by decreasing their expected pickup times. This combination of idle and pickup time creates a non-monotonic U-shaped wait time function, where total wait time is initially decreasing in the number of drivers, then reaches an interior minimum, and then increases in n beyond the minimum.

Driver payoffs are characterized as $u_i = -W_i(n_i)$, so drivers will choose a region where they have the lowest expected wait time. Drivers thus balance idle time and pickup time to determine which market to operate in.

Before laying out definitions of equilibria and turning to our results, we would like to re-iterate our modeling assumptions. In deciding on what assumptions to make, we faced a trade-off between being comprehensive and being able to deliver strong comparative-static results that describe, at the most granular level, how supply redistributes itself spatially in response to a changed market thickness. As such, we decided to abstract away from prices, platform competition, and the dynamic nature of driver behavior. We believe such modeling decisions are supported by the empirical and anecdotal evidence shown in the previous section. Even under these assumptions, proving the

comparative static results is quite involved and requires developing new techniques. Also note that on the issue of modeling the *spatial aspects* of the market, which is of first order relevance to our empirical results, our theory model is in fact more general than the literature: we study a multi-region model in which each region has a size rather than being a point. This is what allows the conceptualization of pickup times and is the main reason why some proofs are involved.

The list of our assumptions follows:

1. Total number of drivers across both markets is fixed at N.
2. Prices are the same for all regions and are, hence, not modeled.
3. Drivers are undifferentiated (conditional on location) and their identity does not matter. Drivers do not have any preference for either of the regions beyond the expected wait times.
4. Each demand arrival gets a location uniformly on the circumference of the circle.
5. The platform greedily allocates consumers to the drivers who are closest to them.
6. The allocation of drivers among regions is thought of as continuous rather than discrete.
7. There is only one platform.

The next subsections define market equilibria and present the results.

5.2 Defining Equilibria and Geographical Supply Inequity

We start by defining what we mean by an equilibrium of this game.

Definition 1. Under "market primitives" (λ, N, t), an allocation $n^* = (n_1^*, ..., n_I^*)$ of drivers among the I regions is called an equilibrium if (i) $\Sigma_{i=1,...,I} n_i^* = N$, and (ii) no driver in any location i can strictly decrease his or her expected total wait time by choosing to drive in another location. Also, we call n^* an "all-regions" equilibrium allocation if it is an equilibrium and if $n_i^* > 0$ for all i.

Next, we define geographical supply inequity.

Definition 2. We say allocation n is under-supplied in region j, relative to region i, if we have:

$$\frac{n_j}{\lambda_j} < \frac{n_i}{\lambda_i}$$

The "degree of under-supply" in region j relative to region i is defined by $\kappa_{ji} = \frac{\frac{n_i}{\lambda_i}}{\frac{n_j}{\lambda_j}}$.

The logic behind this definition is the same as what we had in the empirical section. It basically compares the ratio of the realized numbers of rides n_i between regions to ratio potential demands λ_i.[23]

[23] Unlike the empirical section which had subscripts *ikd*, this section has only i due to the single platform and one-shot nature of the game.

5.3 Results for a Market with Two Regions

In this section, we present our results for the case of $I = 2$. We do this to ease the discussion of the intuition behind our results (since the 2-regions case accepts a simple graphical representation) and to build toward our main theorem. We will present two important results. First, if the demand arrival rate in region 1 is strictly larger than that of region 2, then in any all-regions equilibrium, region 2 will be strictly under-supplied. Second, we show that the under-supply problem in region 2 is mitigated as the size of the platform increases, holding fixed the ratio between λ_1 and λ_2.

First we give a result that helps to visually understand an all-regions equilibrium.

Proposition 1. *At any all-regions equilibrium, the wait times in the two regions are equal. Also the wait time for each region is locally increasing in the number of drivers present in that region.*

Proof. If $W_1(n_1) \neq W_1(n_1)$, then, given the wait time functions are continuous, a small mass of drivers can relocate from the region with the higher wait time to the region with the lower wait time and be strictly better off. Thus, at equilirium allocation n^*, we have $W_1(n_1^*) \neq W_1(n_2^*)$. Next, if at equilibrium, the wait time curve in region i is strictly decreasing, then a small mass of drivers from region j can relocate to i and become strictly better off.■

Next, we introduce a result that speaks to the existence and uniqueness of an all-regions equilibrium.

Proposition 2. *There is exactly one all-regions equilibrium if assumptions (A1) to (A3) hold. Otherwise, there is no all-regions equilibrium.*

(A1) $N \geq \sqrt{\lambda_1 t} + \sqrt{\lambda_2 t}$

(A2,A3) $2\sqrt{\frac{t}{\lambda_j}} \leq \frac{N - \sqrt{\lambda_j t}}{\lambda_i} + \frac{t}{N - \sqrt{\lambda_j t}}$ *for* $j = 1, 2$ *and* $i = 3 - j$

Figure 8 visually illustrates Propositions 1 and 2. In each panel, the wait time curves for the two regions are plotted opposite from each other. In each region, the wait time is initially decreasing in the total number of drivers present in that region due to the decrease it causes in pickup times. But as the region gets more drivers, the effect on pickup time dwindles and overall wait time increases due to increased idle time for drivers.[24] Each point on the horizontal axis of the graph corresponds to a driver allocation between the two regions. One such point is the "demand-proportional" allocation which satisfies $\frac{n_1}{\lambda_1} = \frac{n_2}{\lambda_2}$. This allocation is shown in the figure by a dashed vertical gray line. At each point, the solid blue line shows the total wait time in region 1, and the dashed green line gives the total wait time in region 2.

[24]Total wait time curves being U-shaped has been mentioned in other studies (such as Castillo et al. (2017)). To our knowledge, this curve and the U-shaped assumption on it are used by ride-share platforms in the determination of various strategies including surge pricing.

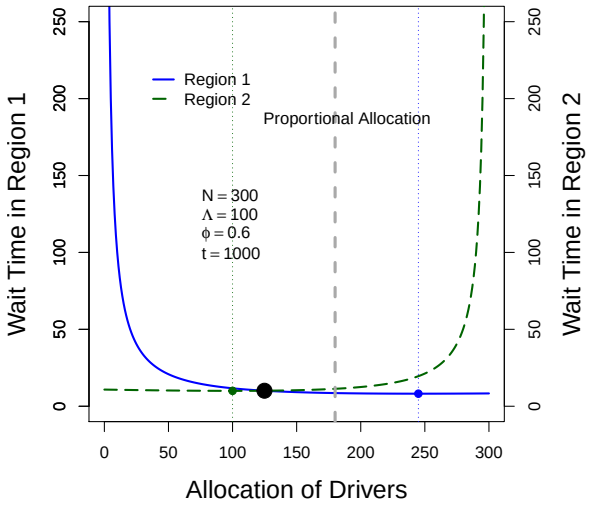
Translated to these graphical terms, Proposition 1 states that an all-regions equilibrium is a point of intersection between the two wait-time curves, at which both curves are increasing. Among the panels of Fig. 8, such equilibrium only exists in panel (c).[25] Proposition 2 explains why. In order for an all-regions equilibrium to exist, there should exist allocations for which the total wait time at each region is increasing in the number of drivers present in that region. This is what assumption **(A1)** requires. Graphically, the trough for the wait time curve in region 2 (emphasized by a green circle) should be to the right of the trough of the wait time in region 1 (blue circle). Panel (a) in Fig. 8 lacks this feature and, hence, also lacks an all-regions equilibrium. In addition to **(A1)**, the existence of an all-regions equilibrium would also require that the two wait-time curves do intersect over the range in which they are both increasing. In order for this to happen, we require assumptions **(A2)** and **(A3)**. They require that, under the allocation that minimizes the total wait time in region 1, the total wait time in region 2 be higher than that in region 1. They impose a similar condition on the allocation that minimizes the total wait time in region 2. Graphically, they require that the total-wait-time curve for region 2 (the green dashed line) be above the trough of the wait-time curve in region 1 (the blue circle), and vice versa. Panel (c) satisfies both **(A2)** and **(A3)** and, hence, has an all-regions equilibrium given by the intersection between the two wait-time curves, emphasized by a large black circle. Panel (b), although satisfying **(A1)**, has the wait time curve for region 1 pass below the trough of the wait time curve in region 2. Therefore, there is no all-regions equilibrium in panel (b).

The reason why different panels in Fig. 8 differ in terms of having an all-regions equilibrium is that they pertain to different market primitives (λ, N, t) (in the figure as well as some of the proofs in the appendix, instead of its components λ_1 and λ_2, the vector λ is represented by total demand $\Lambda = \lambda_1 + \lambda_2$ and share of region 1 from demand $\phi = \frac{\lambda_1}{\Lambda}$). The figures are already suggestive of what affects the existence of an all-regions equilibrium (e.g., a large enough N is necessary) or where the all-regions equilibrium is located when it exists (to the right of the gray dashed line –i.e., the demand-proportional allocation– instead of on it due to agglomeration of drivers in region 1). Our next results in this section formalize and generalize such observations from the figure and add other results describing the role of market thickness.

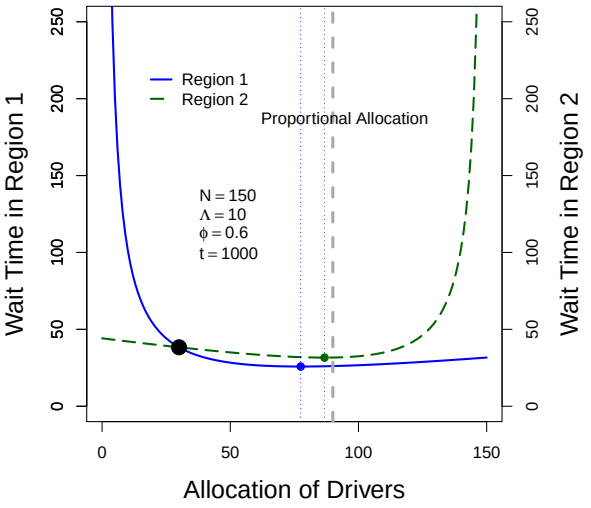
Proposition 3. *Suppose that $\lambda_1 > \lambda_2$ and that an all-regions equilibrium (n_1^*, n_2^*) exists. In that*

[25] One can verify that in all panels of Fig. 8, allocations that put all drivers in one of the two regions are in fact equilibria. To illustrate why, note that under allocation $(n_1, n_2) = (N, 0)$, the wait time at region 2 is ∞ due to high pickup time. Thus, no driver has an incentive to move from region 1 to region 2. This, of course, is an artefact of our assumption of a continuous mass of drivers: If we assume that one driver has a non-trivial mass, some of these one-region equilibria will go away. That said, because of the convenience of assuming a continuous mass of drivers and to avoid multiple equilibria, the rest of the paper focuses on all-regions equilibria only. However, interesting results can be obtained on the properties of the other equilibria and on how their existence and form respond to market thicnkess. All those results, which would be available upon request, have similar economic implications to the results presented on all-regions equilibria.

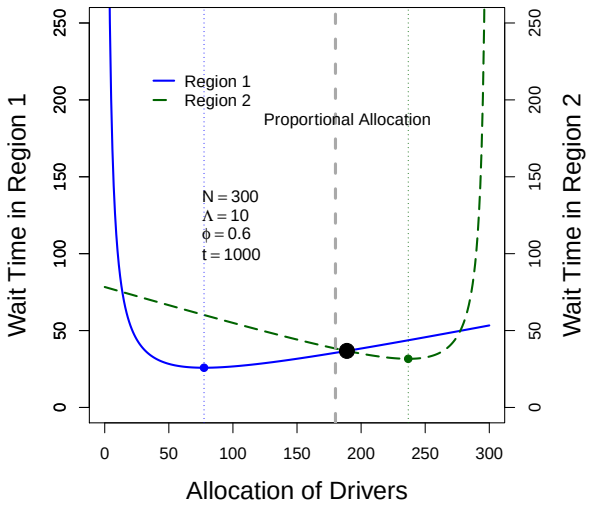
Figure 8: Wait Time and Driver Allocation. An all-regions equilibrium exists only in panel (c)



(a) No All-Regions Equilibrium



(b) No All-Regions Equilibrium



(c) Unique All-Regions Equilibrium

case, the all-regions equilibrium is strictly under-supplied in region 2:

$$\frac{n_1^*}{\lambda_1} > \frac{n_2^*}{\lambda_2}$$

To illustrate, if region 1 has 80% of the demand, then, in equilibrium, 90% of the drivers might prefer to drive in region 1. This result coincides with our empirical observation that the relative outflow was greater in busier areas than in less busy areas.

The proofs for all propositions are given in the appendix. The basic intuition for this proposition is rather simple. Consider an allocation with no under-supply in either region. That is, an allocation with $\frac{n_1}{\lambda_1} = \frac{n_2}{\lambda_2}$. Based on the expressions for idle and pickup times in those areas, it is immediate that under such allocation, the idle times in the two regions are equal, whereas the pickup time is higher in region 2. Therefore, it would be natural to expect drivers to prefer to relocate to region 1, pushing the equilibrium in a direction in which region 2 will be under-supplied.[26] This can be graphically seen in Fig. 8 panel (c): the equilibrium is to the right of the gray dashed line representing the proportional allocation.

We now turn to our second result which speaks to the impact of market thickness. We prove that "making the market thicker" will decrease the extent of geographical inequity in supply. The next two results show this, respectively, for thickening the market on both sides (increasing the number of drivers and all-regional demand arrival rates) and thickening it on one side (increasing the number of drivers only).

Proposition 4. *Suppose that $\lambda_1 > \lambda_2$, and (n_1^*, n_2^*) is the all-regions EQ under $(\lambda_1, \lambda_2, N, t)$. Consider scaling up the platform size by $\gamma > 1$ to $(\lambda_1', \lambda_2', N', t) = (\gamma\lambda_1, \gamma\lambda_2, \gamma N, t)$. Under these new primitives, an all-regions equilibrium exists and under-supply in region 2 decreases with scaling up, i.e., $\frac{n_1'^*}{N'} < \frac{n_1^*}{N}$ In particular, as $\gamma \to \infty$,the relative under-supply in region 2, κ_{21}, tends to zero.*

This proposition speaks to the impact of market thickness (platform size) on geographical supply inequity in two ways. First, it shows that a scale-up in size preserves the existence of an all-regions equilibrium. This means it is possible that as size scales down, drivers abandon a region all-together, making all-regions equilibrium cease to exist. However, as the size scales up, an all-regions equilibrium always remains in existence.

Second, and more importantly, Proposition 4 shows that the all-regions equilibrium under a thicker market exhibits less geographical inequity. To illustrate this result, it says that if under $(\lambda_1, \lambda_2, N, t)$ region 1 had 80% of the total demand, but n_1^* was 90% of N, then under the scaled-up setting $(\lambda_1', \lambda_2', N', t)$, region 1 still has 80% of the total demand but will get, say, 85% of the total number of drivers. The theorem also says that if the size undergoes an extreme scale-up, then region 1 will get very close to 80% of the total number of drivers in the all-regions equilibrium.

Proposition (4) is also proved in the appendix. The intuition behind this proof is that as size gets larger and larger, the platform will get denser in both regions, reducing the importance of pickup times compared to idle times in the decision-making processes of drivers. To show this, we first observe that the extent of geographical supply inequity κ_{21} in the equilibrium is invariant to

[26] The actual proof requires more than this simple intuition. Specifically, it requires a lemma that shows if (A1) through (A3) hold, then each total-wait-time curve is increasing at the demand-proportional allocation (or, put graphically, the troughs of the two curves are located on different sides of the gray dashed line representing the demand-proportional allocation). See appendix for a lemma proving this argument as well as for the details on why such a lemma helps prove the proposition.

multiplying all primitives (i.e., $\lambda_1, \lambda_2, N, t$) by the same factor γ. Therefore, the effect of multiplying only λ_1, λ_2, N by some $\gamma > 1$ on geographical supply inequity will be the same as that of dividing t by γ and holding λ_1, λ_2, N fixed. A division of t by $\gamma > 1$ means drivers care less about pickup times. Drivers caring less about pickup times leads the equilibrium allocation to be closer to what would be implied by idle times only. It is easy to verify that if it were only the idle time that mattered to drivers, the equilibrium allocation would always be one that involved no under-supply in either region: $(n_1^*, n_2^*) = (\frac{N\lambda_1}{\lambda_1+\lambda_2}, \frac{N\lambda_2}{\lambda_1+\lambda_2})$. This is exactly what will be the case as the scale-up grows infinitely large.

Our next proposition proves similar results to those shown in Proposition 4, but this time for thickening the market only on one side.

Proposition 5. *Suppose that $\lambda_1 > \lambda_2$ and (n_1^*, n_2^*) is the all-regions EQ under $(\lambda_1, \lambda_2, N, t)$. If we scale up to $(\lambda_1, \lambda_2, N', t)$ for some $N' > N$, then an all-regions equilibrium still exists. Also, the new equilibrium shows less under-supply of rides in region 2. In particular, as $N' \to \infty$, under-supply in region 2 (and in region 1) tends to zero.*

The proof for this proposition is given in the appendix. The intuition is as follows: a scale-up in N to $N' = \gamma N$ for some $\gamma > 1$ can be thought of as a combination of two changes. First, a scale-up from $(\lambda_1, \lambda_2, N, t)$ to $(\gamma\lambda_1, \gamma\lambda_2, \gamma N, t)$. Second, a scale back down in the demand arrival rates from $(\gamma\lambda_1, \gamma\lambda_2)$ to (λ_1, λ_2). The first move is guaranteed to mitigate the geographical supply inequity problem, according to Proposition (4). The second move increases the importance of idle times (relative to pickup times) in drivers' decision-making processes. Therefore, this change also shifts the new equilibrium toward what would be implied by idle times only, which would be an allocation with no geographical supply inequity.

5.4 Main Result

Our main result extends all of the results presented so far from two regions to any number of regions $I \geq 2$. This theorem is powerful in that it provides, among other results, a description of how the market responds to a changed thickness, at the most granular level. That is, it describes what happens to the supply ratio between *any* two regions i, j. As formalized below, the proposition shows that the market responds to a "global thinning" by further agglomerating the supply at the thickest "local markets."

Theorem 1. In the general version of the game (i.e., $I \geq 2$), the following statements are true:

1. For an all-regions equilibrium, the total wait time is equal across all I regions. Also, at the equilibrium allocation, the total-wait-time curve for any region is strictly increasing in the number of drivers present in that region.

2. Any all-regions equilibrium $n^* = (n_1^*, ..., n_I^*)$ is unique.

3. At any all-regions equilibrium, for any $i < j$, we have $\frac{n_i^*}{\lambda_i} \geq \frac{n_j^*}{\lambda_j}$. The inequality is strict if and only if $\lambda_i > \lambda_j$.

4. Suppose an all regions equilibrium $n^* = (n_1^*, ..., n_I^*)$ exists under primitives (λ, N, t) where $\lambda = (\lambda_1, ..., \lambda_I)$. Then, if supply and demand both scale up, that is, under new primitives $(\gamma\lambda, \gamma N, t)$ with $\gamma > 1$, we have:

   - An all-regions equilibrium $n^{*'} = (n_1^{*'}, ..., n_I^{*'})$ exists.
   - The new equilibrium $n^{*'}$ shows less geographical supply inequity than n^* in the sense that for any $i < j$, we have $1 \leq \frac{\frac{n_i^{*'}}{\lambda_i}}{\frac{n_j^{*'}}{\lambda_j}} \leq \frac{\frac{n_i^*}{\lambda_i}}{\frac{n_j^*}{\lambda_j}}$. Both inequalities are strict if and only if $\lambda_i > \lambda_j$.
   - All $\frac{\frac{n_i^{*'}}{\lambda_i}}{\frac{n_j^{*'}}{\lambda_j}}$ tend to 1 as $\gamma \to \infty$

5. The same statement is true if instead of proportionally scaling up both λ and N, we scale up only N.

These results are closely in line with our empirical results from Table 3 and Table 4. Statement 3 above corresponds to the positive coefficient on borough population density (interpretable only under assumptions 1 and 2). Also, statements 4 and 5 are closely in line with the negative coefficient on the interaction of borough population density and platform size (interpretable under assumption 1 plus either of 2 or 3).

The proof of this result can be found in the appendix. It is based on strong induction. The basis of the induction (i.e., the case of $I = 2$) is given by propositions (1) through (5). The induction works in an interrelated way. That is, for instance, in order to show that item 3 from Theorem (1) holds for some $I = I_0 > 2$, we need not only assume that item 3 holds for all $I \in \{2, ..., I_0 - 1\}$, but also that all of the other items of the proposition hold for all $I \in \{2, ..., I_0 - 1\}$. We believe the proof techniques developed in the implementation of this induction (see appendix) can be useful beyond this paper, in the theoretical analysis of geographical demand-supply mismatch in spatial markets.[27]

[27] We would also like to note, without entering the details, that the proof involves more than a straightforward application of the induction. To illustrate this, consider the case of $I = 3$. Suppose the equilibrium allocation under primitives (λ, N, t) is $n^* = (n_1^*, n_2^*, n_3^*)$. Also suppose that once we scale up both N and λ to obtain primitives $(\gamma\lambda, \gamma N, t)$, we have the equilibrium $n^{*'} = (n_1^{*'}, n_2^{*'}, n_3^{*'})$. Assume, under this new equilibrium, that $n_3^{*'} > \gamma n_3^*$. That is, the least dense region is gaining drivers above and beyond the scale-up, as expected. This implies that regions 1 and 2 will, together, have strictly *fewer* drivers than $\gamma(n_1^* + n_2^*)$. But this renders the application of the induction to the set of regions 1 and 2 insufficient, since now those regions have undergone (i) a scale-up of γ in both demand arrival rates and total number of drivers, followed by (ii) loss of some drivers to region 3. According to our previous results, the first change reduces geographical supply inequity between regions 1 and 2, whereas the second change

5.5 Discussion

Before turning to policy implications of the model, we would like to re-emphasize why our theory results on geographic inequity of supply and the role of market thickness are important beyond ridesharing. We first discuss how the notion of geographical inequity relates to efficiency and then describe how crucial the role of thickness is in spatial markets other than ridesharing (such as taxis).

Geographical Inequity and Efficiency. The main purpose of our model was to analyze geographical inequity in supply and its response to market thickness. The model was not developed with the goal of studying efficiency. However, it can still illuminate some (though not all) of the efficiency implications of inequity. Proposition 6 formalizes this.

Proposition 6. *Consider primitives (λ, N, t) with $\lambda_1 > \lambda_I$ and the set $\mathcal{N}$ of all driver allocations defined as $\{n \in \mathbb{R}^I : \Sigma_i n_i = N, \forall i \ n_i > 0\}$. Suppose $n^0 \in \mathcal{N}$ is the "demand-proportional" allocation: $\forall i, j : \kappa^0_{ji} \equiv \frac{\frac{n^0_i}{\lambda_i}}{\frac{n^0_j}{\lambda_j}} = 1$. Also suppose that $n^1, n^2 \in \mathcal{N}$ are such that $\forall i < j : \kappa^2_{ji} \geq \kappa^1_{ji} \geq 1$. That is, both n^1 and n^2 exhibit geographical inequity in supply in favor of higher demand areas, and the inequity is larger under n^2 than under n^1. Then, the following hold:*

1. *Under all allocations $n \in \mathcal{N}$, the average pickup time for drivers is constant at $\frac{I \times t}{N}$.*
2. *Among all $n \in \mathcal{N}$, the allocation n^0 is the unique minimizer of the average driver idle time. Specifically, an all-regions equilibrium allocation has a higher average idle time than n^0.*
3. *The average driver idle time is higher under n^2 compared to n^1.*

Proposition 6 is proved in the appendix. It describes one reason why geographical inequity in supply is inefficient: By choosing a busy region to minimize her own total wait time, a driver leaves a larger negative externality on the market by substantially lengthening the average idle time in the region she joins and the average pickup time in the regions she avoids. As Proposition 6 shows, any equilibrium allocation, compared to the demand-proportional allocation, makes the total idle time worse without improving the total pickup time. Of course our model is one-shot and, by construction, the total number of given rides is constant at N irrespective of the allocation. But in the real ride-share market, there is repetition. Therefore, in reality, inefficiently high total wait times due to agglomeration of drivers can lead to inefficiently low number of rides given in rideshare (and other transportation) markets.

Aside from the argument above, there is (in our view) a more important reason why geographical inequity of supply may be inefficient that our model does not fully capture. That reason is the very

increases it. Thus, by plain application of induction, one cannot show that geographical supply side inequity between regions 1 and 2 decreases at the end. However, we prove lemmas in the appendix which guarantee the proof of the proposition, in spite of the fact that induction applies in some but not all of the cases.

notion of inequity. If residents of region j consistently have lower access to supply of transportation services than region i (i.e., if a higher fraction of the potential demand is forgone in j than in i), then the marginal demand in j is likely to be for more essential transportation needs than in i. This may imply some allocative inefficiency. The issue of inequity has been major topic in the transportation science literature.[28] It has also been salient enough in public policy to bring about such major actions as the launch of green taxis (also called "boro taxis"[29]) However, quantifying the magnitude of the welfare effects of inequity is beyond the scope of our paper. It will require panel data on passengers in order to capture the fact that some groups are regularly under-supplied relative to others.

Generalizability of Results Beyond Ridesharing. Our theoretical model assumes a central dispatch structure for the matching of drivers to riders, and one of the key forces behind our results is the pickup time. Also, all of our empirical analysis is performed on rideshare data. This raises the question of whether the geographical inequity in supply (due to agglomeration) can arise in a market with decentralized matching, such as the taxicab market. In that market, there might be search frictions; but once a cab and a passenger find each other, they are not far apart.

We believe our insights are also crucial in understanding the spatial distribution of supply in the taxicab market. Table 7 corroborates this by presenting the relative outflows for the Yellow Taxis across boroughs of NYC during January 2009, the first month on which data on the taxicab market is available from the TLC. We find it interesting that, similar to the rideshare market, the relative outflows in the taxicab market almost have a perfect rank-correlation with the borough population densities (the only exception is Queens, most likely because areas closer to the airports become denser and, hence, more attractive). In fact, relative outflows are much more skewed toward Manhattan for taxicabs than they are in the rideshare market. For instance, it is interesting to observe that although the outflow of rides from Manhattan was about 742 times more than that from Staten Island, the ratio between the inflows was only 10 (the total population of Manhattan is about 4.2 times that of Staten Island).[30]

Studies that examine the taxicab market in NYC tend to focus on Manhattan, on the grounds that the large majority of rides take place there (see Buchholz (2018); Lagos (2003) for instance). However, based on the above observation, we argue that this is likely an equilibrium outcome in which supply gets highly agglomerated in Manhattan. Therefore, understanding *why* there is such a sharp contrast between Manhattan and the outer boroughs may be of first order importance in

[28] See Litman (1999); Delbosc and Currie (2011); Pereira et al. (2017) among may other references.

[29] For more details, see the history of boro taxis on the TLC website from this link.

[30] The skewness of these relative outflows towards Manhattan would be even more sriking once we notice that in the Taxicab market, as opposed to rideshare, drivers have full discretion on which rides to give. Therefore, drivers in Manhattan might refuse to give rides that exit the borough because they anticipate they will have to return to Manhattan empty. Thus, we conjecture that if it were not for such discretion, the relative outflows for the taxicabs would be even more skewed.

Table 7: Relative Outflows in NYC Boroughs for Yellow Taxi during Jan. 2009

Borough	Outflow	Inflow	Relative Outflow
Bronx	9,436	56,981	0.17
Brooklyn	95,727	406,111	0.24
Manhattan	682,159	161,049	4.24
Queens	72,601	221,218	0.33
Staten Island	919	15,483	0.06

Note: rides to and from airports (i.e., JFK and LGA in Queens) have been excluded.

studying what shapes the geographical distribution of supply in spatial markets with decentralized matching.[31] This is particularly important because the same agglomeration mechanism that leads to the sharp observed contrast *between* Manhattan and other boroughs might also be at work in determining how drivers position themselves *within* Manhattan.[32]

6 Implications for Policy

Our work is timely since it relates to the policy debate on whether rideshare platforms should be downsized. New York has recently been considering implementing multiple policies which, either directly or indirectly, will shrink the size of rideshare platforms. This policy debate is important both because NYC is the largest city in the country and because of the precedent the action taken by NYC will likely set for other cities. One proposed policy is imposing a $17 minimum hourly wage on the rideshare platforms (The Washington Post, 2018; Wired, 2019), which took effect in the beginning of February 2019 (The Hill, 2019). Another policy is to impose a cap on the number of licenses each platform can hand out to drivers (hence a cap on the number of drivers who can drive for these platforms). The particular way this regulation was designed was by halting, for 12 months starting August 2018, the issuance of new licenses for drivers of rideshare platforms (The Verge, 2018; Tech Crunch, 2018). The reactions of ridesharing platforms to the aforementioned regulations

[31] One could think of a theory of agglomeration in de-centralized transportation markets that is similar in nature to the theory in our paper. In the taxicab (rideshare) market, lower density of demand and supply in outer boroughs leads to higher local search frictions (longer pickup times). This geographical difference in search efficiencies (pickup times) in turn distorts the supply further away from the outer boroughs. In fact, Frechette et al. (2019) already document that there is economy of scale in search efficiency, which can lead to overall more efficient search when the market is thicker. It would not be unnatural to think, then, that "where" the market is thicker is more desirable for drivers, hence the self-reinforcing loop.

[32] Indeed, we carried out a relative outflows analysis on the set of taxicab rides from January 2009 that started and ended in Manhattan. We found a sharp contrast in relative outflows between Lower and Central Manhattan (which are where the density of rides are the highest) on the one hand, and Upper Manhattan on the other.

(and potential regulations) have been mostly negative.[33] Finally, a third approach considered by the city is to start levying a "congestion tax" on drivers. The fares for rides originating in lower Manhattan were supposed to increase by $2.50 for taxi and $2.75 for rideshare, effective January 1, 2019. However, the implementation has been temporarily postponed due to a lawsuit brought by a coalition of drivers and taxi owners, calling the tax a "suicide charge" (The New York Times, 2019b).[34] Whether this regulation will eventually be implemented is still uncertain (The New York Times, 2019a).

In this section, we discuss what we can learn from the theoretical and empirical analyses conducted in this paper for public policy issues. We focus on the potential impacts of such policies on the distribution of drivers across the city and on the geographical (in)equity of the availability of rideshare services. Of course, this by no means is a claim that geographical inequity is the only important implication of this policy. For instance, our paper does not focus on the labor-market consequences of this policy nor does it focus on the impact on congestion. Nevertheless, we believe it does bring up an issue for consideration that is important in navigating future decisions.

Some policy tools might have an advantage over others from the perspective of reducing (or not increasing) geographical inequity. For instance, imposing a congestion tax (currently planned to take effect in 2020) might be preferred over downsizing the total number of drivers. Of course, if a congestion tax leads to downsizing rideshare platforms, it will, according to our results, also provide an incentive for drivers to drive in busier areas. However, the tax will provide a direct incentive for drivers to serve less busy areas. Such a "counter-incentive" is not provided by a plain downsizing regulation. In fact, our results could be used to defend a congestion tax policy against the potential criticism that a congestion tax might cause under-supply in busier areas. Our results suggest that downsizing rideshare platforms via a congestion tax leads to driver incentives in both directions (i.e., both to drive less in busier areas and to drive more in less busy areas), whereas a direct downsize of the number of drivers (or a geography-independent mandatory wage increase) would only increase the incentive to drive more in busy areas and less in other areas, exacerbating the inequality problem.

[33] Uber has sued the city of New York over the year-long pause to issuing new ridesharing licenses (The Tech Crunch, 2019). Their spokesperson has claimed that such policy will do little to help mitigate the congestion in NYC (Tech Crunch, 2018). The spokesperson stated that he believed the congestion tax to be a more effective policy regarding controlling congestion. On the equity front, ridesharing platforms contend that a downsize of ridesharing will hit the outer boroughs harder than Manhattan, given that those areas might have lower access to public transportation options and taxis and thereby be more reliant on ridesharing (Tech Crunch, 2018). Also, on the front of fairness among ridesharing platforms, smaller platforms have brought lawsuits against the city for multiple aspects of its crackdown on ridesharing. Lyft and Juno sued the city for the minimum wage regulation which is calculated on a weekly basis, rather than based on hours driven with a passenger. They claimed this hurts smaller platforms with lower utilization rates (Wired, 2019).

[34] The term originates from multiple recent cases of driver suicides in NYC due to financial hardship and the belief that some recent regulations by the city have exacerbated the drivers' situation (The New York Times, 2018).

Another qualitative takeaway from the analysis is that competition policy is complicated by driver location choice. That is, a hypothetical breakup of a large ridesharing firm into two smaller ones could have opposing effects. On the one hand, the competition between the two could benefit consumers. On the other hand, in each of those two smaller platforms (and hence, overall), the under-supply in less busy areas will increase.[35]

On the quantitative side, we answer an interesting question motivated by our theoretical analysis. Our theoretical results show that geographical inequity diminishes as platform size becomes infinitely large due to the fact that pickup times lose their importance against idle times. In a sense, this implies that if the platform size is "large enough," then geographical inequity will not be a first order concern. A practical question is how large is this "large enough" size? To find out, we modify regression equation (7), replacing the log function applied to platform size by a function that satiates to an upper limit as the platform size increases.

We implement this by using $\log(\min(a_{Max}, \#Rides))$ instead of $\log(\#Rides)$, where a_{Max} is the parameter capturing the adequate size and is to be estimated (one could interpret a_{Max} as the size at which the impact of size on the geo-distribution of relative outflows becomes small enough so that it cannot be distinguished from noise). We choose this way of capturing the adequate size over adopting a functional form that converges smoothly as size grows. The reason behind this choice is that we want the identification of the adequate size to come mainly from the data points at which relative outflows stop responding to platform size, as opposed to the data points at which the platform size is well below the upper limit. The regression equation implementing this notion is very similar to the earlier regression Eq. (7) on relative outflows, with the difference being the inclusion of a_{Max}. Equation (10) describes this regression:

$$RO_{ikd} = \alpha_0 + \alpha_1 \log(\rho_i) + \alpha_2 \log(\min(a_{Max}, S_{kd})) + \alpha_3 \log(\min(a_{Max}, S_{kd})) \log(\rho_i) + \nu_{ikd} \quad (10)$$

In order to make sure that the functional form of log is not substantially impacting our estimate of a_{Max}, we also estimate a version in which the size itself, as opposed to its natural log, is used. Equation (11) represents this:

$$RO_{ikd} = \alpha_0 + \alpha_1 \log(\rho_i) + \alpha_2 \min(a_{Max}, S_{kd}) + \alpha_3 \min(a_{Max}, S_{kd}) \log(\rho_i) + \nu_{ikd} \quad (11)$$

[35] It might seem at first that "multi-homing" (i.e., the phenomenon of drivers working for multiple platforms (Bryan and Gans, 2019)) might mitigate the excess clustering of supply of small platforms in busier areas, because drivers working for multiple platforms are, in effect, working for one large rideshare system. We note, however, that, for multi-homing to mitigate agglomeration, it must be that the matching systems across platforms are fully integrated. This would imply, for instance, that a Lyft driver would not get asked to pick up a passenger who is far away, if there is an Uber driver in the vicinity of that passenger. We believe that in reality, the integration of matching systems is substantially less than perfect, rendering multi-homing less impactful on the extent of agglomeration. Indeed, if multi-homing could eliminate agglomeration, it should have shown up in the relative outflows of Lyft and Via in Fig. 4.

Regressions (10) and (11) are estimated using non-linear least squares, and the results are reported in Table (8). The adequate size parameter, a_{Max} is estimated at 3.65M rides/month using regression (10) and at 3.30M rides/month using regression (11). Both estimates are statistically very significant. They are also fairly close to each other, suggesting the robustness of a_{Max} to the model specification, as we expected.

These results suggest that NYC needs to use caution if it were to downsize Lyft and, especially, Via (see numbers reported in Fig. 4). Uber, on the other hand will not face distorted geographical supply distribution if downsized. We note that given a similar dataset to what we used here, the method we laid out in this section can help identify a_{Max} in any other metropolitan area.[36]

7 Conclusion

This paper asked three questions about the functioning of spatial markets and studied them in the context of the rideshare market in NYC: (i) How can we empirically identify whether there is geographical demand-supply mismatch, leaving some regions with persistently lower access to supply compared to others? (ii) What mechanism leads to such persistent geographical inequity in supply? (iii) How should we design policies that help mitigate the inequity? To answer these questions, we started by developing the "relative-outflows" method. It is fairly simple to implement, has limited data requirements, detects under-supply in a region even if passengers in that region have, over the long run, learned not to search for rides, and finally can be applied to markets with centralized or decentralized matching in the same way. We used this method to show that rideshare platforms (especially smaller ones) tend to be under-supplied in low-population-density regions. As such, we conducted an empirical study pointing to the role of market thickness (platform size) on the geographical balance between demand and supply. We complemented it with a theory model that studies the impact of market thickness on the geographical distribution of supply. We showed that making the market thinner skews the supply ratio between any two regions toward the higher density one, even though demand ratios are fixed. Finally, on the policy front, we estimated a minimum required size for rideshare platforms in NYC in order to avoid the overclustering of supply in busier areas. Our method could be used to find such required sizes in other metropolitan

[36] We consider these estimates of a_{Max} to be lower bounds in the sense that the minimum required size may be larger than they are. The reason is, even at Uber's current size, Uber's relative outflows are skewed toward Manhattan in terms of magnitude (though less so than the other platforms). Under assumptions (1) and (3), our empirical method does not allow us to identify whether this is because of under-supply of Uber in the outer boroughs or because of geographical heterogeneity in outside options (because our method can only identify cross-platform differences). But under assumptions (1) and (2), even Uber's current size would be too small, making our estimated a_{Max} a lower bound. But even then, we believe this estimate is very useful because it shows at what size the response of the geographical distribution of supply to size becomes so slow that even with an almost three-fold growth in size (from Lyft to Uber), only a negligible improvement in geographical equity of supply is achieved.

Table 8: Results of regressions (10) and (11)

	Dependent variable: Relative Outflow	
	(1)	(2)
Regression	Equation (10)	Equation (11)
α_0	-15.07^{***}	-1.449^{***}
	(0.2829)	(0.027)
α_1	4.129^{***}	6.408^{***}
	(0.079)	(7.604e-03)
α_2	1.030^{***}	$5.876\text{e-}07^{***}$
	(0.019)	(1.254e-08)
α_3	-0.264^{***}	$-1.505\text{e-}07^{***}$
	(0.005)	(3.428e-09)
a_{Max}	$3.648\text{e+}06^{***}$	$3.295\text{e+}06^{***}$
	(7.120e+04)	(3.916e+04)
Observations	7,709	7,709
Note:		*p<0.1; **p<0.05; ***p<0.01

The main coefficient of interest is a_{Max}, the adequate size for a rideshare platform to contain geographical inequity in supply.

areas as well.

Our research can be extended along a number of dimensions. On the theory side, one interesting question would be about the role of platform incentives and pricing. More specifically, it is not entirely clear whether a platform should "go along" with its drivers agglomerating in busier areas or whether it should try to "correct" the agglomeration. On the one hand, the time spent by drivers on the way to pick a passenger up is a loss both to them and to the platform, suggesting that the drivers' action to avoid long pickup times by relocating to busy areas is in line with what the platform would want. On the other hand, a driver's decision to relocate to another area impacts not only his or her wait times, but also other drivers' wait times. In particular, it may increase the pickup time in the quieter area more than it decreases the pickup time in the busy area. What this suggests is that the platform might want to intervene and mitigate agglomeration through prices. A theoretical model, more general than the one we built in this paper, is needed to address this question and characterize the optimal intervention by the platform.

On the empirical side, quantifying the consumer welfare effects of the geographical inequity in supply would be a major step. We believe a prerequisite to such a study would be panel data on passengers in order to capture the fact that persistent under-supply of rides in a region means persistent under-supply of rides to the same population, which could have large adverse effects if the marginal utility from taking a ride diminishes with the number of rides taken. Another interesting direction for future research would be to empirically study whether the impact of agglomeration on the spatial distribution of drivers across a city is comparable to or larger than that of other mechanisms studied in the literature. For instance, Lagos (2000, 2003) focus on the role of the average length of rides starting in each region on the attractiveness of that region for drivers. Buchholz (2018) focuses on how drivers' decisions are impacted by the inter-temporal, intra-daily, externalities from rides given by other drivers. Brancaccio et al. (2019c) study the inefficiency arising from transportation of goods/passengers to locations from which the car/ship would likely need to return vacant. The modeling of all (or even a subset) of the above, in conjunction with spatial network externalities that lead to agglomeration, could substantially complicate the computation and/or make the estimation of the parameters too reliant on parametric assumptions. This presents the difficult question of what (not) to include in empirical models of such comoplex markets. As such, we believe an extension of our study, from a sole examination of agglomeration to an empirical comparison between magnitude of agglomeration and those of the other forces mentioned above, can enrich the literature.

References

Afèche, P., Liu, Z., and Maglaras, C. (2018). Ride-hailing networks with strategic drivers: The impact of platform control capabilities on performance.

Ahlfeldt, G. M., Redding, S. J., Sturm, D. M., and Wolf, N. (2015). The economics of density: Evidence from the berlin wall. *Econometrica*, 83(6):2127–2189.

Akbarpour, M., Li, S., and Gharan, S. O. (2017). Thickness and information in dynamic matching markets.

Ashlagi, I., Burq, M., Jaillet, P., and Manshadi, V. (2019). On matching and thickness in heterogeneous dynamic markets. *Operations Research*.

Banerjee, S., Kanoria, Y., and Qian, P. (2018). The value of state dependent control in ridesharing systems. *arXiv preprint arXiv:1803.04959*.

Bimpikis, K., Candogan, O., and Saban, D. (2016). Spatial pricing in ride-sharing networks. *Available at SSRN 2868080*.

Brancaccio, G., Kalouptsidi, M., and Papageorgiou, T. (2019a). Geography, transportation, and endogenous trade costs. *Econometrica, Forthcoming*.

Brancaccio, G., Kalouptsidi, M., and Papageorgiou, T. (2019b). A guide to estimating matching functions in spatial models.

Brancaccio, G., Kalouptsidi, M., Papageorgiou, T., and Rosaia, N. (2019c). Efficiency in decentralized transport markets.

Bryan, K. A. and Gans, J. S. (2019). A theory of multihoming in rideshare competition. *Journal of Economics & Management Strategy*, 28(1):89–96.

Buchholz, N. (2018). Spatial equilibrium, search frictions and dynamic efficiency in the taxi industry. Technical report, Working Paper.

Buchholz, N., Shum, M., and Xu, H. (2018). Dynamic labor supply of taxicab drivers: a semiparametric optimal stopping model. *Available at SSRN 2748697*.

Cachon, G. P., Daniels, K. M., and Lobel, R. (2017). The role of surge pricing on a service platform with self-scheduling capacity. *Manufacturing & Service Operations Management*, 19(3):368–384.

Castillo, J. C., Knoepfle, D., and Weyl, G. (2017). Surge pricing solves the wild goose chase. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 241–242. ACM.

Castro, F., Besbes, O., and Lobel, I. (2018). Surge pricing and its spatial supply response.

Chen, M. K., Chevalier, J. A., Rossi, P. E., and Oehlsen, E. (2017). The value of flexible work: Evidence from uber drivers. Technical report, National Bureau of Economic Research.

Chen, M. K. and Sheldon, M. (2016). Dynamic pricing in a labor market: Surge pricing and flexible work on the uber platform. In *Ec*, page 455.

Cohen, P., Hahn, R., Hall, J., Levitt, S., and Metcalfe, R. (2016). Using big data to estimate consumer surplus: The case of uber. Technical report, National Bureau of Economic Research.

Cramer, J. and Krueger, A. B. (2016). Disruptive change in the taxi business: The case of uber. *American Economic Review*, 106(5):177–82.

Datta, S. and Sudhir, K. (2011). The agglomeration-differentiation tradeoff in spatial location choice. *manuscript. Yale School of Management*.

Delbosc, A. and Currie, G. (2011). Using lorenz curves to assess public transport equity. *Journal of Transport Geography*, 19(6):1252–1259.

Ellison, G. and Glaeser, E. L. (1997). Geographic concentration in us manufacturing industries: a dartboard approach. *Journal of political economy*, 105(5):889–927.

Frechette, G. R., Lizzeri, A., and Salz, T. (2019). Frictions in a competitive, regulated market: Evidence from taxis. Technical Report 8.

Guda, H. and Subramanian, U. (2019). Your uber is arriving: Managing on-demand workers through surge pricing, forecast communication, and worker incentives. *Management Science*.

Holmes, T. J. (2011). The diffusion of wal-mart and economies of density. *Econometrica*, 79(1):253–302.

Lagos, R. (2000). An alternative approach to search frictions. *Journal of Political Economy*, 108(5):851–873.

Lagos, R. (2003). An analysis of the market for taxicab rides in new york city. *International Economic Review*, 44(2):423–434.

Lam, C. T. and Liu, M. (2017). Demand and consumer surplus in the on-demand economy: the case of ride sharing. *Social Science Electronic Publishing*, 17(8):376–388.

Lian, Z. and van Ryzin, G. (2019). Optimal growth in two-sided markets. *Available at SSRN 3310559*.

Litman, T. (1999). *Evaluating transportation equity*. Victoria Transport Policy Institute Victoria, BC, Canada.

Miyauchi, Y. (2018). Matching and agglomeration: Theory and evidence from japanese firm-to-firm trade. Technical report.

Nikzad, A. (2018). Thickness and competition in ride-sharing markets. Technical report, Working paper, Stanford University.

Pereira, R. H., Schwanen, T., and Banister, D. (2017). Distributive justice and equity in transportation. *Transport reviews*, 37(2):170–191.

Petrongolo, B. and Pissarides, C. A. (2001). Looking into the black box: A survey of the matching function. *Journal of Economic literature*, 39(2):390–431.

Shapiro, M. H. (2018). Density of demand and the benefit of uber.

Tech Crunch (2018). New york city council votes to cap licenses for ride-hailing services like uber and lyft.

The Hill (2019). Uber competitors say demand has plunged under new nyc rules.

The New York Times (2018). Why are taxi drivers in new york killing themselves?

The New York Times (2019a). Congestion pricing plan for manhattan ran into politics. politics won.

The New York Times (2019b). suicide surcharge or crucial fee to fix the subway? taxi drivers brace for battle over 2.50*charge*.

The Tech Crunch (2019). Uber sues nyc to contest cap on driversrules.

The Verge (2018). In major defeat for uber and lyft, new york city votes to limit ride-hailing cars.

The Washington Post (2018). New rules guarantee minimum wage for nyc uber, lyft drivers.

Wired (2019). Lyft sues new york city over driver minimum wage rules.

Appendices

A Anecdotal Evidence from Media and Online Forums

This appendix points to a list of anecdotal pieces of evidence (by no means exhaustive) from online rideshare forums on how drivers complain about Lyft's far pickups in suburbs and how they recommend responding to it. The explanations in brackets withing the quotations are from us.

- **From the online forum "Uber People,"[37] a thread in the Chicago section:** The title of the thread is "To those who drive Lyft in the suburbs." The thread was started on Dec 19 2016. The first post says "Are the ride requests you get on Lyft always seem to be far away from you location? Seems like they are always 5 miles or more for the pickup location. I got one for 12 miles last night. I drive in the Schaumburg/Palatine area [two northwestern suburbs of Chicago about 30mi away from downtown]."

[37] In spite of what the name suggests, this is a general ridesharing forum, not exclusively about Uber.

- **From the same thread:** "iDrive primarily in Palatine. about two out of every five ride requests are for more than 10 minutes away. I ignore those".
- **From the same thread:** "I was a victim of that once. Never again I take a ping more than 10 minutes away in the burbs".
- **From the same thread:** "Yesterday was my 1st day on Lyft. Was visiting in Homer Glen [a village about 30mi southwest of downtown Chicago] & decided to try Lyft for the first time. First ping was 18 minutes away. Dang, I could make it 1/2 way downtown in that time! I ignored the ride request. 2nd ping was also 18 minutes away. Lyft app complained my acceptance rate is too low. I ignored the 2nd ping & went off-line."
- **From the same forum, a thread titled "First 3days of Lyft":** "If your area is spread out...and you have to take those ¿ 10 minute requests, well...I might look for another job."
- **From the same thread:** "Yeah, another (mostly) Lyft-specific problem, especially when working in the suburbs, is you sometimes (fairly frequently, actually) receive trip requests that are not close to your current location. I've received requests from passengers 20 miles away."
- **From Chicago Tribune article titled "Lyft takes on Uber in suburbs":** Jean-Paul Biondi, Chicago marketing lead for Lyft is quoted to explain the reason for Lyft's planned expansion into surburbs as follows "The main reason is we saw a lot of dropoffs in those areas, but people couldn't get picked up in those areas." Which is in line with our reasoning that small relative-outflow is a sign of potential demand which does not get served due to under-supply.
- **From the rideshare website "Become a Rideshare Driver":** It says successful Lyft drivers use the following strategy:
  - "The drivers usually run the Lyft app exclusively when they are in the busy downtown or city areas."
  - "Usually in the suburbs, Uber is busier than Lyft, and in such areas, the drivers run both the Uber and Lyft apps."

B Hourly Analysis of Relative and Absolute Flows

In this section, we test Assumption 3 in the empirical part of the paper. As mentioned in Section 4, we assume that the attractiveness of the outside option to riders can change in a platform-specific or direction-specific way but *not in a platform-direction specific way*. Such exclusion restriction cannot be directly tested but we offer one indirect test:

Assumption 3 will be violated if (i) users of platform k tend to need exit rides from region i at systematically different times of the day than when users of platform k' exit the region; and (ii) the availability of outside transportation options changes with time of day. To illustrate, suppose Uber users in Staten Island are systematically more likely than Lyft users to need a ride exiting the borough in early-morning hours when public transit is less available. Under this violation of Assumption 3, the relative outflow of Uber in Staten Island will be larger than that of Lyft due to the demand-side differences among users, *not due to supply-side differences* as we claimed in the main text of the paper.

To examine whether such violation of Assumption 3 is likely, we can compare different platforms' relative outflows over differnt hours of the day. If most of the differences between the relative

outflows of two platforms can be explained by looking at certain times of the day, then we should be concerned about the validity of Assumption 3. If however, the comparison between relative outflows of platforms is consistent throughout the day, we would be more confident about Assumption 3. Figures 9 and 10 plot the absolute and relative flows of different platforms in Staten Island for each hour of the day, averaged over days of July 2017 or June 2018. As can be seen there, during July 2017, Lyft's relative outflow is consistently lower than Ubers throughout the day. During June 2018, Lyft's and Uber's relative (and absolute) flows move together consistently, and Via's is consistently and substantially lower than both of them. We find similar patterns when analyzing other boroughs. This gives us more confidence that Assumption 3 is reasonable, and, hence, the differences in relative outflows across platforms can be attributed to the supply side.

C Proofs

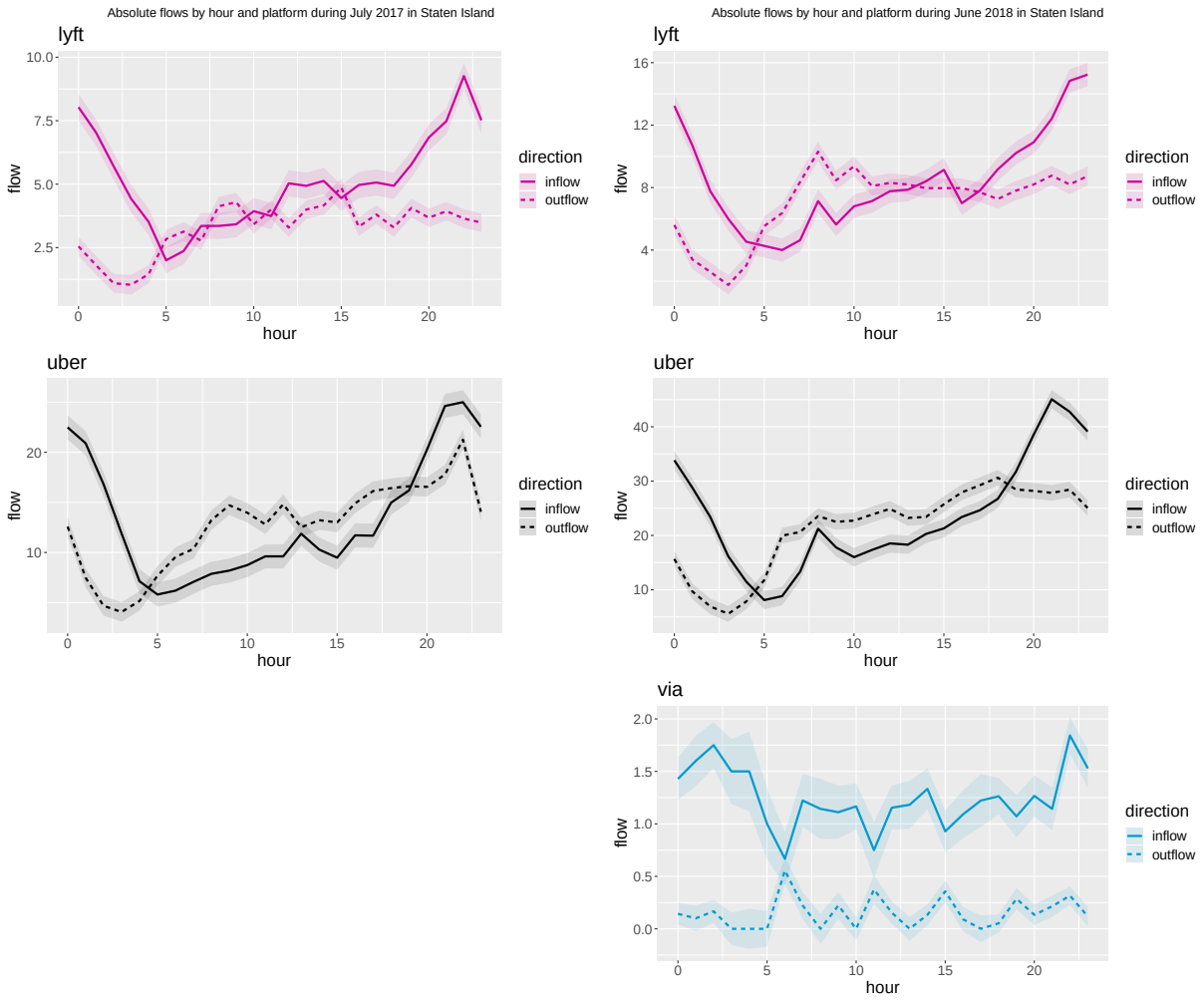
Proof of Proposition 2. First, we prove the necessity part, and then sufficiency and uniqueness.

Necessity: We prove necessity of (A1)-(A3) by contradiction. First, if (A1) is not satisfied, then we show that the wait time curves can only intersect when at least one of them is decreasing. To see this, note that taking the first order condition on eq. (9) shows the wait time curve in each region i is minimized at $n_i^{min} = \sqrt{\lambda_i t}$. Thus, condition (A1) simply requires that $N \geq n_1^{min} + n_2^{min}$. Without (A1), there would be no possible allocation of drivers under which the total wait time in each region is increasing in the number of drivers present in that region. Therefore, by Proposition 1, there would be no all-regions equilibrium. Next, suppose condition (A2) were not true. Thus, at the minimum wait time for region 1, i.e. at the allocation $(n_1 = n_1^{min}, n_2 = N - n_1^{min})$, the wait time for region 1 is higher than region 2. Thus, the wait time curves can only intersect in the decreasing region for market 1, which we know cannot be an all-regions equilibrium. The necessity of (A3) is similar to (A2).

Sufficiency: Observe that when (A2) is true, $W_1(n_1^{min}) < W_2(N - n_1^{min})$. Similarly from (A3), we have $W_2(n_2^{min}) < W_1(N - n_2^{min})$. We know that for $n_1 > n_1^{min}$, W_1 is an increasing function, and similar is the case for W_2. Since we have a reversal in relative magnitude for W_1 and W_2, and since the two curves are continuous, we must have an intersection of the curves between n_1^{min} and n_2^{min}, when both wait time curves are increasing. Such an intersection permits no profitable deviation by switching to the other market for any driver, and is thus an all-regions equilibrium.

Uniqueness: Both wait time curves are monotonic for the region $n_1 > n_1^{min}$ and $n_2 = (N - n_1) > n_2^{min}$, implying that there can only be one intersection between the curves when they are both increasing.

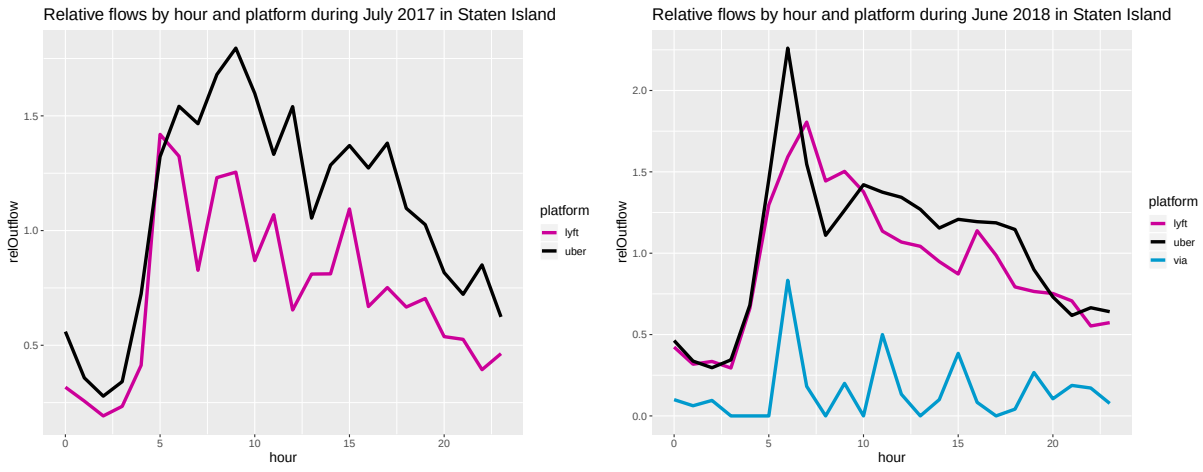
Figure 9: Absolute inflows outflows for Lyft, Uber, and Via† in Staten Island (hourly averages), the shadowed areas show the 95% confidence intervals.



(a) (b)

†: Panel (a) is July 2017 and Panel (b) is June 2018

Figure 10: Relative outflows for Lyft, Uber, and Via† in Staten Island (hourly averages)



(a) (b)

†: Panel (a) is July 2017 and Panel (b) is June 2018

Together, these conditions are proven equivalent to existence and uniqueness of an all-regions equilibrium. In such a case, we can characterize the all-regions equilibrium by equating the wait time distributions.[38] ■

To Prove Proposition 3, we first introduce the following Lemma.

Lemma A1. *When (A1)-(A3) are satisfied and when drivers are allocated proportionally to demand, the proportional allocation lies between the minimum wait times for the two regions:* $n_1^{min} < \phi N < N - n_2^{min}$.

where ϕ, as mentioned in the main text, is defined as $\phi = \frac{\lambda_1}{\lambda_1+\lambda_2}$. By our assumption $\lambda_1 > \lambda_2$, which came without loss of generality, we have $\phi > \frac{1}{2}$. In graphical terms represented by Fig. 8, this lemma says the vertical dashed line representing the proportional demand will fall between the troughs of the two wait-time curves.

Proof of Lemma A1. *First, we prove that the proportional allocation line lies between the two minima.* $n_1^{min} = \sqrt{\lambda_1 t}$ *and* $n_2^{min} = \sqrt{\lambda_2 t}$. *Denote the total demand across both locations as* $\Lambda = \lambda_1 + \lambda_2$ *and the fraction of demand in the (higher-demand) location* 1 *to be* $\phi = \frac{\lambda_1}{\Lambda} > \frac{1}{2}$.

[38] In practice, we obtain the allocation equating wait times, i.e. solving $W_1(n) - W_2(N-n) = 0$, which is equivalent to identifying the roots of the polynomial equation below:

$$-n^3(\lambda_1+\lambda_2) + n^2(2N\lambda_1 + N\lambda_2) - n(N^2\lambda_1 + 2t\lambda_1\lambda_2) + Nt\lambda_1\lambda_2 = 0$$

By Descartes' rule of signs, this equation (i.e. the numerator) has potentially 3 positive roots. In the case of multiple roots, only the one that lies between the minimum points of the wait time curves where both curves are increasing is the symmetric equilibrium. See Proposition 1.

For proportional allocation to be situated between the two minimums on the graph, the following conditions need to hold:

C1: $\phi N > n_1^{min} = \sqrt{\lambda_1 t} = \sqrt{\phi \Lambda t} \implies N > \sqrt{\frac{\Lambda t}{\phi}}$

C2: $(1-\phi)N > n_2^{min} = \sqrt{\lambda_2 t} = \sqrt{(1-\phi)\Lambda t} \implies N > \sqrt{\frac{\Lambda t}{1-\phi}}$

Observe that since $\phi > \frac{1}{2}$, $C2 \implies C1$. Thus, when the demand is more skewed (higher ϕ), we need to have a larger platform size for condition (C2) to be satisfied.

We now prove that assumption (A1) + (A3) $\implies$ (C2). Observe that condition that shows up is the following: Assumption (A3) implies

$$2\sqrt{\frac{t}{(1-\phi)\Lambda}} < \frac{N - \sqrt{(1-\phi)\Lambda t}}{\phi\Lambda} + \frac{t}{N - \sqrt{(1-\phi)\Lambda t}}$$

The second term on the RHS can be bounded as: $\frac{t}{N-\sqrt{(1-\phi)\Lambda t}} < \frac{t}{\sqrt{\phi\Lambda t}}$, since $N > \sqrt{\phi\Lambda t} + \sqrt{(1-\phi)\Lambda t}$ by assumption (A1).

Thus assumption (A3) implies the following:

$$\frac{N - \sqrt{(1-\phi)\Lambda t}}{\phi\Lambda} > 2\sqrt{\frac{t}{(1-\phi)\Lambda}} - \frac{t}{\sqrt{\phi\Lambda t}} \implies N > \sqrt{\Lambda t}\left(2\phi\sqrt{\frac{1}{1-\phi}} - \sqrt{\phi} + \sqrt{1-\phi}\right)$$

Next, we prove that the above inequality implies condition (C2), which stated that $N > \sqrt{\frac{\Lambda t}{1-\phi}}$. Thus, we need to prove the following:

$$2\phi\sqrt{\frac{1}{1-\phi}} - \sqrt{\phi} + \sqrt{1-\phi} > \sqrt{\frac{1}{1-\phi}} \Leftrightarrow \frac{2\phi - 1}{\sqrt{1-\phi}} - \sqrt{\phi} + \sqrt{1-\phi} > 0$$
$$\Leftrightarrow \sqrt{\phi}(\sqrt{\phi} - \sqrt{1-\phi}) > 0$$

Observe that the last inequality must be true given our assumption that $\phi > \frac{1}{2}$, so (A1) + (A3) $\implies$ (C2). This finishes the proof of the lemma. ■

Proof of Proposition 3. At proportional allocation, by Lemma A1, the demand-proportional allocation is in between the minimum wait times for both regions. We show the proportional allocation or any point to the left of it (i.e., an alloction with $n_1 \leq \phi N$) cannot be an all-regions equilibrium. This, combined with the assumption in the proposition that an all-regions equilibrium exists, implies that the all-regions equilibrium should be to the right of the proportional allocation. That is: $\frac{n_1^*}{\lambda_1} > \frac{n_2^*}{\lambda_2}$.

To see why no all-regions equilibrium can be found weakly to the left of the proportional allocation, note that at proportional allocation $n = \phi N$, region 2 wait time is higher than region 1, i.e. $W_2((1-\phi)N) = \frac{N}{\Lambda} + \frac{t}{(1-\phi)N} > W_1(\phi N) = \frac{N}{\Lambda} + \frac{t}{(\phi)N}$ since $\phi > \frac{1}{2}$. As we move left, market 2's wait time increases further, while market 1's wait time decreases until we reach the minimum

wait time for market 1, $W(n_1^{min})$. Thus, the divergence between the two markets increases. For the wait time curves to intersect, it must be in market 1's decreasing wait time region. We know from Proposition 1 that such an intersection will **not** be an all-regions equilibrium. Fig. 8 panel (c) should help illustrate this point. This completes the proof of the proposition. ■

Lemma A2. *When an all-regions equilibrium exists for a ridesharing platform with N drivers facing demand $\phi\Lambda$ and $(1-\phi)\Lambda$ in the two regions:*

1. *an all-regions equilibrium also exists when demand is unchanged and there are $N' = \gamma N$ drivers where $\gamma > 1$.*

2. *an all-regions equilibrium also exists when both the demand and number of drivers is scaled by $\gamma > 1$ to $N' = \gamma N$ and $\Lambda' = \gamma\Lambda$.*

Proof of Lemma A2. *Consider the equivalent conditions required for the existence of an all-regions equilibrium, characterized by assumptions (A1)-(A3). Below, we show that if the conditions are satisfied for a given (N, Λ), then they must be satisfied for (a) $(N', \Lambda') = (\gamma N, \Lambda)$ as well as (b) $(N', \Lambda') = (\gamma N, \gamma\Lambda)$.*

First, consider (A1). The proof of (a) is immediate. For (b), we observe that:

$$\gamma N > \sqrt{\gamma}\sqrt{\Lambda t}\left(\sqrt{\phi} + \sqrt{1-\phi}\right)$$

holds since $\gamma > 1$ and (A1) holds for (N, Λ).

Next, we prove (A2). The proof of (A3) is similar to that of (A2) and is omitted.

For (A2), first we denote the following function ϕ:

$$\psi(\rho) = \frac{\rho N - \sqrt{\phi\Lambda t}}{(1-\phi)\Lambda} + \frac{t}{\rho N - \phi\Lambda t}$$

We prove that ψ is increasing in ρ, or $\frac{d\phi}{d\rho} > 0$. Observe that:

$$\frac{d\psi}{d\rho} = N\left(\frac{1}{(1-\phi)\Lambda} - \frac{t}{\left(\rho N - \sqrt{\phi\Lambda t}\right)^2}\right)$$

After some algebra and applying (A1), we obtain $\frac{d\psi}{d\rho} > 0$.

Now, for part (a), observe that setting $\rho = \frac{N'}{N} > 1$ implies that, in (A2), the RHS increases and the LHS does not change implying that (A2) still holds for $(N', \Lambda') = (kN, \Lambda)$.

Next, for (b), observe that applying (A2) with $(N', \Lambda') = (\gamma N, \gamma\Lambda)$ gives us:

$$2\sqrt{\frac{t}{\gamma\phi\Lambda}} < \frac{\gamma N - \sqrt{\phi\gamma\Lambda t}}{(1-\phi)\gamma\Lambda} + \frac{t}{\gamma N - \phi\gamma\Lambda t} \Leftrightarrow 2\sqrt{\frac{t}{\phi\Lambda}} < \frac{\sqrt{\gamma}N - \sqrt{\phi\Lambda t}}{(1-\phi)\Lambda} + \frac{t}{\sqrt{\gamma}N - \phi\Lambda t}$$

We need to prove the above holds whenever (A1)-(A3) hold.

Since we know that ψ is an increasing function, we know that $\psi(\sqrt{\gamma}) > \psi(1)$ when $\gamma > 1$. But if we write out $\psi(\sqrt{\gamma}) > \psi(1)$, it gives us exactly the expression we needed to be true:

$$2\sqrt{\frac{t}{\phi\Lambda}} < \frac{\sqrt{\gamma}N - \sqrt{\phi\Lambda t}}{(1-\phi)\Lambda} + \frac{t}{\sqrt{\gamma}N - \phi\Lambda t}$$

Thus, when $(N', \Lambda') = (\gamma N, \gamma\Lambda)$, we find that (A2) holds for (N', Λ').

Thus, (A1)-(A3) hold under the conditions detailed in the Lemma. ■

Proof of Proposition 4. The proof of existence of all-regions equilibrium under the new model primitives obtains from Lemma A2 above. To prove that the equilibrium supply ratios tilts towards region 2, we first claim (but skip the straightforward proof) that if all the primitives of the model (λ, N, t) are multiplied by same scaling factor, the existence of an all-regions equilibrium as well as all of the $\frac{n_i^*}{n_j^*}$ ratios (and, by construction, all $\frac{\lambda_i}{\lambda_j}$ ratios) are preserved. Therefore, in this proof, instead of a multiplication of N and λ by a factor of $\gamma > 1$, we focus on fixing N and λ and, instead, replacing t by $t\frac{1}{\gamma}$.

For the all-regions equilibrium (n_1^*, n_2^*), define $\alpha = \frac{n_1^*}{N}$. We know from Proposition 3 that $\alpha > \phi$. The quilibrium condition, written in terms of α will be:

$$W^d(\alpha) \equiv W_1(\alpha N) - W_2((1-\alpha)N) = -\frac{(1-\alpha)N}{(1-\phi)\Lambda} + \frac{(\alpha)N}{\phi\Lambda} + \frac{t}{(1-\alpha)N} + \frac{t}{(\alpha)N} = 0 \tag{12}$$

where W^d represents the difference between the total wait times between the two regions, which should be zero at the equilibrium. We now use the implicit function theorem to show that α increases as we increase t, which would prove the proposition.

$$\frac{d\alpha}{dt} = -\frac{\frac{\partial W^d}{\partial t}}{\frac{\partial W^d}{\partial \alpha}} = \frac{(1-\alpha)\alpha(2\alpha-1)(1-\phi)\phi\Lambda}{(\alpha-1)^2\alpha^2N^2 + (2(\alpha-1)\alpha+1)(\phi-1)\phi\Lambda t} \tag{13}$$

The numerator is positive since $\alpha > \frac{1}{2}$. Thus, the sign of $\frac{d\alpha}{dt}$ is determined by the denominator. Below, we prove that the denominator is positive as well. The argument takes the following steps:

1. Define the denominator as $g(\alpha) = (\alpha-1)^2\alpha^2N^2 + (2(\alpha-1)\alpha+1)(\phi-1)\phi\Lambda t$.
2. Observe that $g'(\alpha) = -2(2\alpha-1)\left((1-\alpha)\alpha N^2 + (1-\phi)\phi\Lambda t\right) < 0$, implying that $g(\alpha)$ is a decreasing function.
3. Since $\alpha \in \left[\phi, 1 - \frac{n_2^{min}}{N}\right]$, the inequality $g(\alpha) \geq g\left(1 - \frac{n_2^{min}}{N}\right)$ has to hold.
4. We prove that $\min g(\alpha) = g\left(1 - \frac{n_2^{min}}{N}\right) > 0$.

$$g\left(1 - \frac{n_2^{min}}{N}\right) = \frac{\phi^2\Lambda t}{N^2}\left(N^2 - 2N\sqrt{\phi\Lambda t} + (2\phi-1)\Lambda t\right) \tag{14}$$

$$= \frac{\phi^2\Lambda t}{N^2}\left((N - \sqrt{\phi\Lambda t})^2 - t\Lambda(1-\phi)\right) \tag{15}$$

where the term in parentheses is positive directly from assumption **(A1)**.

Thus, we know that $\frac{d\alpha}{dt} > 0$ implying that as t increases, the proportion of supply going to the higher-demand market is greater. ■

Proof of Proposition 5. As mentioned in the text of the paper, a scale-up in N can be thought of as a scale-up in $(\lambda_1, \lambda_2, N)$, followed by a scale back down in (λ_1, λ_2). From Proposition 4, we know that the first scale-up (i) preserves the existence of an all-regions equilibrium and also (ii) makes it strictly less under-supplied in region 2. Therefore, the proof of Proposition (5) will be complete if we show that the second scale back down also preserves the existence of an all-regions equilibrium and makes it less under-supplied inn region 2.

To see this, suppose (n_1^*, n_2^*) is the all-regions equilibrium under $(\lambda_1, \lambda_2, N, t)$. Let $\lambda_i' = \frac{\lambda_i}{\gamma}$ for $i \in \{1, 2\}$ and some $\gamma > 1$. We will now show that under $(\lambda_1', \lambda_2', N, t)$, there is an all-regions equilibrium with strictly less under-supply in region 2 than what is implied by (n_1^*, n_2^*).

Lemma A3. *The following statements are true about the "old" equilibrium allocation (n_1^*, n_2^*) under the "new" parameters $(\lambda_1', \lambda_2', N, t)$:*

1. *The total wait function $W_2(n)$ is strictly increasing at $n = n_2^*$.*
2. *At the allocation (n_1^*, n_2^*), the wait time in region 1 is strictly higher than that in region 2. That is, $W_1(n_1^*) > W_2(n_2^*)$.*
3. *The total wait function $W_1(n)$ is strictly increasing at $n = N \times \frac{\lambda_1'}{\lambda_1' + \lambda_2'}$.*
4. *At the allocation proportional to demand, the wait time in region 2 is strictly larger than that in region 1. That is, if we set $n_i = N \times \frac{\lambda_i'}{\lambda_1' + \lambda_2'}$, then $W_2(n_2) > W_1(n_1)$.*

Proof of Lemma A3. We start by statement 1. To see this, first note that from the assumption that (n_1^*, n_2^*) was the all-regions equilibrium under $(\lambda_1, \lambda_2, N, t)$, we know n_2^* has to be strictly larger than where the old W_2 function reached its trough. That is, $n_2^* > \sqrt{\lambda_2 t}$. Now, given $\lambda_2' < \lambda_2$, we it is also the case that $n_2^* > \sqrt{\lambda_2' t}$. Therefore, the new W_2 function is also strictly increasing at $n = n_2^*$.

We now turn to statement 2. Given that (n_1^*, n_2^*) was the all-regions equilibrium under the old parameters, the total wait times in the two regions were equal to each other. That is:

$$\frac{n_1^*}{\lambda_1} + \frac{t}{n_1^*} = \frac{n_2^*}{\lambda_2} + \frac{t}{n_2^*} \tag{16}$$

Given Proposition (3), we know that $n_1^* > n_2^*$, therefore: $\frac{t}{n_1^*} < \frac{t}{n_2^*}$. This latter inequality, combined with equality (16), implies $\frac{n_1^*}{\lambda_1} > \frac{n_2^*}{\lambda_2}$. The sign of this inequality is preserved if we multiply both sides of it by the positive number $\gamma - 1$. That is: $(\gamma - 1) \times \frac{n_1^*}{\lambda_1} > (\gamma - 1) \times \frac{n_2^*}{\lambda_2}$.

The size of the inequality is also preserved when we add equal numbers to both sides. Those equal numbers are the two sides of equation (16). This will give us:

$$(\gamma-1)\times\frac{n_1^*}{\lambda_1}+\frac{n_1^*}{\lambda_1}+\frac{t}{n_1^*}>(\gamma-1)\times\frac{n_2^*}{\lambda_2}+\frac{n_2^*}{\lambda_2}+\frac{t}{n_2^*} \tag{17}$$

Therefore:

$$\gamma\times\frac{n_1^*}{\lambda_1}+\frac{t}{n_1^*}>\gamma\times\frac{n_2^*}{\lambda_2}+\frac{t}{n_2^*} \tag{18}$$

which gives us:

$$\frac{n_1^*}{\frac{\lambda_1}{\gamma}}+\frac{t}{n_1^*}>\frac{n_2^*}{\frac{\lambda_2}{\gamma}}+\frac{t}{n_2^*} \tag{19}$$

which, by definition, means:

$$\frac{n_1^*}{\lambda_1'}+\frac{t}{n_1^*}>\frac{n_2^*}{\lambda_2'}+\frac{t}{n_2^*} \tag{20}$$

This proves statement 2.

Next, we turn to statement 3. The argument is similar to that for statement 1. $N\times\frac{\lambda_1}{\lambda_1+\lambda_2}$ was larger than the trough of W_1 under the old parameters. Given that the trough gets smaller under the new parameters, it will keep being smaller than $N\times\frac{\lambda_1}{\lambda_1+\lambda_2}$.

Finally, statement 4 is obvious from the proof of Proposition (3). □

Now notice that Lemma (A3) completes the proof of the proposition. Given that under the allocation (n_1^*,n_2^*), we have $W_1>W_2$, and under the allocation fully proportional to demand, we have $W_1<W_2$, and given that both W_1 and W_2 are continuous functions, there should be an allocation $(n_1^{*'},n_1^{*2})$ in between the two such that $W_1(n_1^{*'})=W_2(n_2^{*'})$. This was achieved by statements 2 and 4. Now by statement 1, W_2 is strictly increasing at $n=n_2^{*'}$ because $n_2^{*'}>n_2^*>\sqrt{\lambda_2't}$. Also, by statement 3, W_1 is increasing at $n=n_1^{*'}$ because $n_1^{*'}>N\frac{\lambda_1}{\lambda_1+\lambda_2}>\sqrt{\lambda_1't}$. This implies that $(n_1^{*'},n_2^{*'})$ is the all-regions equilibrium under parameters $(\lambda_1',\lambda_2',N,t)$. Now, given $n_1^{*'}<n_1^*$ and $n_2^{*'}>n_2^*$, it follows that:

$$\kappa^{*\prime}=\frac{\frac{n_1^{*'}}{\lambda_1'}}{\frac{n_2^{*'}}{\lambda_2'}}<\frac{\frac{n_1^*}{\lambda_1'}}{\frac{n_2^*}{\lambda_2'}}=\frac{\frac{n_1^*}{\lambda_1}}{\frac{n_2^*}{\lambda_2}}=\kappa^*$$

which finishes the proof of the proposition.■

Proof of Theorem (1). Before stating the induction hypothesis, we add one statement to the five statements of Theorem (1). The inclusion of this statement and leveraging it in the induction process will be helpful for the proof. We call it statement 6.

Statement 6. Suppose an all region equilibrium $n^* = (n_1^*, ..., n_I^*)$ exists under primitives (λ, N, t) where $\lambda = (\lambda_1, ..., \lambda_I)$. Then, demand arrival rates are scaled down, that is, under new primitives $(\frac{\lambda}{\gamma}, N, t)$ with $\gamma > 1$, we have:

- *An all-regions equilibrium $n^{*'} = (n_1^{*'}, ..., n_I^{*'})$ exists.*
- *The new equilibrium $n^{*'}$ shows less geographical supply inequity than n^* in the sense that for any $i < j$, we have $\frac{\frac{n_i^{*'}}{\lambda_i}}{\frac{n_j^{*'}}{\lambda_j}} \leq \frac{\frac{n_i^*}{\lambda_i}}{\frac{n_j^*}{\lambda_j}}$. The inequality is strict if and only if $\lambda_i > \lambda_j$.*

In words, this statement simply says the geographical supply inequity decreases if, all else fixed, all demand arrival rates proportionally decrease. The intuition is that this makes idle times relatively more important than pickup times.

We can now state the strong induction hypothesis.

Induction Hypothesis. Take some natural number $I_0 > 2$. If all statements of Theorem (1), including statement 6 added above, are correct for $I \in \{2, ..., I_0 - 1\}$, then they are also all correct for $I = I_0$.

Now, in order to prove the theorem, we need to take two steps. First, we should prove the basis of the induction process. That is, we must show the theorem holds under $I = 2$. Second, we need to prove the induction hypothesis. As for the first step, note that propositions (1) through (5) do this job. The only statement that is not explicitly proven by those theorem is statement 6. However, the proof of statement 6 was the main building block of the proof of Proposition (5).[39]

We now turn to the second and main step of this proof, which is to show that the induction hypothesis is correct (Note that some of the statements are not really proven based on the induction. Nevertheless, we present all of the proofs in this inductive framework since we believe having one induction as well as one non-induction section for the proof will just make it harder to read).

Proof of Statement 1. If the total wait time in region i is strictly higher than that in region j, then given the continuity of these wait-time functions, a small enough mass of drivers can leave region i for j and strictly benefit from that, violating the equilibrium assumption. To see why they are increasing, suppose on the contrary, that at the equilibrium allocation, for region i, the total wait time is strictly decreasing in the number of drivers in that region. Since drivers are equal across all regions in equilibrium, drivers from any other region j will have the incentive to relocate to region i, given that (i) currently region i has the same total wait as they do; and (ii) once they move to region i, the total wait time of that region will decrease. This is a violation of the equilibrium assumption. Therefore it has to be that at the equilibrium, the wait times are all increasing in the number of drivers at all regions.□

[39] Also, propositions (1) through (5) assume that $\lambda_1 > \lambda_2$ and, hence, leave out the case where $\lambda_1 = \lambda_2$. But the proofs for the case where $I = 2$ and $\lambda_1 = \lambda_2$ are straightforward and we leave them to the reader.

Proof of Statement 2. Suppose, on the contrary, that there are two different all-regions equilibria n^* and $\bar{n}$. Given the two vectors are different, there has to be a region i such that $n_i^* \neq \bar{n}_i$. Without loss of generality, assume $n_i^* < \bar{n}_i$. Given that, from statement 1, we know the total wait time is increasing at n_i^*, and given the fact that the wait time function, once it becomes increasing, it remains strictly increasing, we can say $W_i(n_i^*) < W_i(\bar{n}_i)$.

Now, again from statement 1, we know two things. First, $\forall j: \quad W_j(n_j^*) = W_i(n_i^*) \quad \& \quad W_j(\bar{n}_j) = W_i(\bar{n}_i)$, which implies: $\forall j: \quad W_j(n_j^*) < W_j(\bar{n}_j)$. Second, we know that the total wait time function at each region j must be strictly increasing after it hits its trough (which happens weakly before n_j^*). This implies that in order for $\forall j: \quad W_j(n_j^*) < W_j(\bar{n}_j)$ to hold, it must be that $\forall j: \quad n_j^* < \bar{n}_j$. Therefore:

$$\Sigma_{j=1,...,I_0} \bar{n}_j > \Sigma_{j=1,...,I_0} n_j^*$$

But this cannot be given that both of the sums should be equal to N.□

Proof of Statement 3. Note that the definition of equilibrium is that no driver should have the incentive to relocate from one region to another. This definition, by construction, implies that if $n^* = (n_1^*, ..., n_{I_0}^*)$ is an equilibrium under (λ, N, t), then once we fix $\tilde{N} = n_i^* + n_j^*$ for some i, j with $i < j$, then the allocation (n_i^*, n_j^*) is itself an equilibrium of the two-region game with primitives $(\lambda_i, \lambda_j, \tilde{N}, t)$. Thus, by Proposition (3) (or alternatively, by the base of the induction), we know that if $\lambda_i > \lambda_j$, then $\frac{n_i^*}{\lambda_i} > \frac{n_j^*}{\lambda_j}$. Also in case $\lambda_i = \lambda_j$, it is fairly straightforward to verify that $\frac{n_i^*}{\lambda_i} = \frac{n_j^*}{\lambda_j}$. To see this, note that in that case, $\frac{n_i^*}{\lambda_i} = \frac{n_j^*}{\lambda_j}$ if and only if $n_i^* = n_j^*$. It is easy to see that $n_i^* = n_j^*$ is an equilibrium given that it gives the two regions the same total wait time and that at it, the total wait times must be increasing according to previous statements.□

Proof of Statement 4. Before we start the proof of this statement, we note that, similar to the case of Proposition (4), we can work with primitives $(\lambda, N, \frac{t}{\gamma})$ instead of $(\gamma\lambda, \gamma N, t)$. As a reminder, this is because there is a one-to-one and onto mapping between the equilibria under the two primitives, which preserves all of the $\frac{n_i^*}{\lambda_i}$ values.

We start by proving the first statement. That is, if an all-regions equilibrium exists under (λ, N, t), then one does under $(\lambda, N, \frac{t}{\gamma})$ as well. To see this, let us assume that under the "old" primitives (λ, N, t), the all-regions equilibrium allocation n^* is such that $\forall i \in \{1, ..., I_0\}: \quad W_i(n_i^*) = w$. We know this common w must exist from statement 1, and we know it is unique from statement 2. We show existence of an equilibrium allocation under the new primitive by first describing two "partial equilibrium" allocations. We construct the first partial equilibrium allocation $\bar{n} = (\bar{n}_1, ..., \bar{n}_{I_0})$ by fixing $\bar{n}_1 = n_1^*$ and assuming the rest of values $(\bar{n}_2, ..., \bar{n}_{I_0}$ to be the equilibrium allocation of drivers among regions 2 to I_0 under primitives $((\lambda_2, ..., \lambda_{I_0}), N - n_1^*, \frac{t}{\gamma})$. In words, this allocation fixes the number of drivers in region 1 (i.e., the region with the highest demand arrival rate λ_1) at its value under the old primitives but allows the drivers of all other regions to reshuffle themselves among those regions. The second partial equilibrium allocation $\tilde{n}$ fixes $\tilde{n}_1 = N \times \frac{\lambda_1}{\Sigma_{i=1,...,I_0} \lambda_i}$, and

assumes the rest of values $(\tilde{n}_2, ..., \tilde{n}_{I_0}$ to be the equilibrium allocation of drivers among regions 2 to I_0 under primitives $((\lambda_2, ..., \lambda_{I_0}), N \times (1 - \frac{\lambda_1}{\Sigma_{i=1,...,I_0}\lambda_i}), \frac{t}{\gamma})$. In words, this allocation fixes the total number of drivers in region 1 at the value it would take if drivers were to be allocated fully proportional to demand arrival rates. It then allows the rest of the drivers to reshuffle themselves among other areas under the new primitives. We will use these two partial equilibrium allocations to prove existence of an all-region equilibrium. But first we need to prove the existence of these partial equilibrium allocations themselves. Lemma A4 below does this job.

Lemma A4. *Partial equilibrium allocations $\tilde{n}$ and $\bar{n}$ described above exist, are unique, and allocate a strictly positive number of drivers to each region.*

Proof of Lemma (A4). We first start from $\bar{n}$. Note that the assumption of n^* being the equilibrium allocation under (λ, N, t), by construction implies that $(n_2^*, ..., n_{I_0}^*)$ is the unique all-region equilibrium allocation under $((\lambda_2, ..., \lambda_{I_0}), N - n_1^*, t)$. Now, given that by our induction assumption all results (including statement 4) hold for $I_0 - 1$ regions, if the primitives remain the same except that t is divided by some $\gamma > 1$, a unique all-region equilibrium will still exist. This is what we were denoting $\bar{n}_2$ through $\bar{n}_{I_0}$.

Next, we turn to $\tilde{n}$ and construct it from $\bar{n}$. We just showed that $(\bar{n}_2, ..., \bar{n}_{I_0})$ is the unique all-regions equilibrium under $((\lambda_2, ..., \lambda_{I_0}), N - n_1^*, t)$. Also note that by statement 3, we know $n_1^* > \frac{\lambda_1}{\Sigma_{i=1,...,I_0}\lambda_i})$, which implies $N - n_1^* < N \times (1 - \frac{\lambda_1}{\Sigma_{i=1,...,I_0}\lambda_i}))$. Therefore, primitives $((\lambda_2, ..., \lambda_{I_0}), N \times (1 - \frac{\lambda_1}{\Sigma_{i=1,...,I_0}\lambda_i}), \frac{t}{\gamma})$ can be constructed from primitives $((\lambda_2, ..., \lambda_{I_0}), N - n_1^*, t)$ by increasing the total number of drivers. Given that $\bar{n}$ was the unique all-regions equilibrium allocation under $((\lambda_2, ..., \lambda_{I_0}), N - n_1^*, t)$, and given the induction assumption on statement 5 for $I = I_0 - 1$ regions, we can say that primitives $((\lambda_2, ..., \lambda_{I_0}), N \times (1 - \frac{\lambda_1}{\Sigma_{i=1,...,I_0}\lambda_i}), \frac{t}{\gamma})$ also have a unique all-regions equilibrium allocation. This is exactly what was denoted $\tilde{n}_2, ..., \tilde{n}_{I_0}$. This completes the proof of the lemma.□

We now use these two partial equilibrium allocations to show that a unique all-regions equilibrium allocation exists under primitives $(\lambda, N, \frac{t}{\gamma})$. Our next step is to prove the following useful lemma.

Lemma A5. *At the partial equilibrium allocation $\bar{n}$, the total wait time in region 1 is larger than that in any other region. Conversely, at the partial equilibrium allocation $\tilde{n}$, the total wait time in region 1 is smaller than that in any other region.*

Proof of Lemma (A5). To see why the result holds for $\bar{n}$, note that under the old equilibrium n^* and old primitives (λ, N, t), all of the wait times were equal. This means for any $i > 1$ we had

$$\frac{n_1^*}{\lambda_1} + \frac{t}{n_1^*} = \frac{n_i^*}{\lambda_i} + \frac{t}{n_i^*}$$

But given that for all $i > 1$ we have $n_1^* \geq n_i^*$ we get the following inequality under the new primitives $(\lambda, N, \frac{t}{\gamma})$:

$$\frac{n_1^*}{\lambda_1} + \frac{\frac{t}{\gamma}}{n_1^*} \geq \frac{n_i^*}{\lambda_i} + \frac{\frac{t}{\gamma}}{n_i^*}$$

Next, note that the main and only difference between allocations n^* and $\bar{n}$ is that under $\bar{n}$, drivers reshuffle among regions 2 to I_0 in order to reduce their total wait times. Therefore, there has to be at least one region j such that:

$$\frac{\bar{n}_j}{\lambda_j} + \frac{\frac{t}{\gamma}}{\bar{n}_j} \leq \frac{n_j^*}{\lambda_j} + \frac{\frac{t}{\gamma}}{n_j^*}$$

Combining the above two, we get:

$$\frac{\bar{n}_j}{\lambda_j} + \frac{\frac{t}{\gamma}}{\bar{n}_j} \leq \frac{n_1^*}{\lambda_j} + \frac{\frac{t}{\gamma}}{n_1^*}$$

But the total wait time under $(\lambda, N, \frac{t}{\gamma})$ is equal across regions 2 through I_0 under allocation $\bar{n}$. Therefore, the above inequality holds not only for a specific j , but under any $j > 1$. This proves the lemma for $\bar{n}$ given that $\bar{n}_1 = n_1^*$.

Next, we prove the lemma for $\tilde{n}$. We first show that the wait time in region 1 is smaller than that in region 2 if both get drivers proportional to their demand arrival rates. We then show that the wait time in region 2 under $\tilde{n}_2$ is larger than the wait time in region 2 if region 2 were to get drivers proportional to its demand arrival rate. These two statements, combined, will prove our intended result. To see the first claim, note that the wait time in region 1, if it gets $N \times \frac{\lambda_1}{\Sigma_{i\in\{1,\ldots,I_0\}}\lambda_i}$ drivers, will be:

$$w_1 = \frac{N \times \frac{\lambda_1}{\Sigma_{i\in\{1,\ldots,I_0\}}\lambda_i}}{\lambda_1} + \frac{\frac{t}{\gamma}}{N \times \frac{\lambda_1}{\Sigma_{i\in\{1,\ldots,I_0\}}\lambda_i}}$$

which gives:

$$w_1 = \frac{N}{\Sigma_{i\in\{1,\ldots,I_0\}}\lambda_i} + \frac{\frac{t}{\gamma} \times \Sigma_{i\in\{1,\ldots,I_0\}}\lambda_i}{N\lambda_1} \tag{21}$$

Similarly, if region 2 were to get $N \times \frac{\lambda_2}{\Sigma_{i\in\{1,\ldots,I_0\}}\lambda_i}$ drivers, its total wait time will be:

$$w_2 = \frac{N}{\Sigma_{i\in\{1,\ldots,I_0\}}\lambda_i} + \frac{\frac{t}{\gamma} \times \Sigma_{i\in\{1,\ldots,I_0\}}\lambda_i}{N\lambda_2} \tag{22}$$

It is easy to see that the first terms of w_1 and w_2 are the same, and the second term is larger in w_2 given that $\lambda_1 \geq \lambda_2$. Now note that under allocation $\tilde{n}$, the wait time in region 1 is indeed w_1. So, it remains to show that $W_2(\tilde{n}_2) \geq w_2$. To show this, we make two observations (and prove them both shortly). First, $\tilde{n}_2 \geq N \times \frac{\lambda_2}{\Sigma_{i\in\{1,\ldots,I_0\}}\lambda_i}$. This simply says under $\tilde{n}$, region 2 is getting more drivers than it would if drivers were to be allocated to regions proportionally to their demand rates. Second, the total wait time function in region 2 is increasing between $N \times \frac{\lambda_2}{\Sigma_{i\in\{1,\ldots,I_0\}}\lambda_i}$ and $\tilde{n}_2$. Together, these two observations imply $W_2(\tilde{n}_2) \geq w_2$, as desired. Therefore, we have shown that $W_2(\tilde{n}_2) \geq w1$.

But given that $(\tilde{n}_2, ..., \tilde{n}_{I_0})$ was an all-regions equilibrium under $((\lambda_2, ..., \lambda_{I_0}), N \times (1 - \frac{\lambda_1}{\Sigma_{i=1,...,I_0}\lambda_i}, \frac{t}{\gamma})$, we know that for any $i, j > 1$: $\quad W_i(\tilde{n}_i) = W_j(\tilde{n}_j)$. This, combined with $W_2(\tilde{n}_2) \geq w_1$, completes the proof of the lemma, of course with the exception of the two observations made in this paragraph. We now turn to proving those observations and finish the proof of the lemma.

The first observation was that $\tilde{n}_2 \geq N \times \frac{\lambda_2}{\Sigma_{i\in\{1,...,I_0\}}\lambda_i}$. To see why this is true, note that $\tilde{n}$ is the all-regions equilibrium under $((\lambda_2, ..., \lambda_{I_0}), N \times (1 - \frac{\lambda_1}{\Sigma_{i=1,...,I_0}\lambda_i}), \frac{t}{\gamma})$. Therefore, by our induction assumption on statement 3, region 2 will get disproportionately more drivers relative to all other regions, because it has the highest λ_i amongst regions $2, ..., I_0$. That is $\forall i > 2 : \quad \frac{\tilde{n}_2}{\lambda_2} \geq \frac{\tilde{n}_i}{\lambda_i}$. It is then easy to show that:

$$\frac{\tilde{n}_2}{\lambda_2} \geq \frac{\Sigma_{i=2,...,I_0}\tilde{n}_i}{\Sigma_{i=2,...,I_0}\lambda_i} \tag{23}$$

But we know, from the primitives, that $\Sigma_{i=2,...,I_0}\tilde{n}_i = N \times (1 - \frac{\lambda_1}{\Sigma_{i=1,...,I_0}\lambda_i}) = N \times \frac{\Sigma_{i=2,...,I_0}\lambda_i}{\Sigma_{i=1,...,I_0}\lambda_i}$. Now, plugging this into (23) and rearranging, we get $\tilde{n}_2 \geq N \times \frac{\lambda_2}{\Sigma_{i\in\{1,...,I_0\}}\lambda_i}$, which is exactly our first observation.

We now turn to the proof of the second observation. That is, we want to show that the total wait time function in region 2 is increasing between $N \times \frac{\lambda_2}{\Sigma_{i\in\{1,...,I_0\}}\lambda_i}$ and $\tilde{n}_2$. To see this, note that the wait time curve in region 2 takes the form that was depicted in figure (8). In particular, it is a curve with only one trough; and once past the trough, the curve will remains strictly increasing indefinitely. Thus, to prove that the wait-time is increasing over the interval $[N \times \frac{\lambda_2}{\Sigma_{i\in\{1,...,I_0\}}\lambda_i}, \tilde{n}_2]$, it is sufficient to show that the smallest point in this interval is past the trough. One can show the trough happens at $n_2 = \sqrt{\frac{t}{\gamma}\lambda_2}$. Therefore, what we need to show is:

$$N \times \frac{\lambda_2}{\Sigma_{i\in\{1,...,I_0\}}\lambda_i} \geq \sqrt{\frac{t}{\gamma}\lambda_2} \tag{24}$$

In order to prove this, we first assume, to the contrary, that $N \times \frac{\lambda_2}{\Sigma_{i\in\{1,...,I_0\}}\lambda_i} < \sqrt{\frac{t}{\gamma}\lambda_2}$; then we arrive at a contradiction with the result that $\tilde{n}$ is the all-regions equilibrium under $((\lambda_2, ..., \lambda_{I_0}), N \times (1 - \frac{\lambda_1}{\Sigma_{i=1,...,I_0}\lambda_i}, \frac{t}{\gamma})$. Note that we are assuming, without loss of generality, $\lambda_2 \geq \lambda_i$ for any $i > 2$. Therefore, given that all λ_i are positive, for any $i > 2$, we have $\frac{\lambda_i}{\lambda_2} \leq \sqrt{\frac{\lambda_i}{\lambda_2}}$. Thus, if we multiply the left hand side of the inequality $N \times \frac{\lambda_2}{\Sigma_{i\in\{1,...,I_0\}}\lambda_i} < \sqrt{\frac{t}{\gamma}\lambda_2}$ by $\frac{\lambda_i}{\lambda_2}$ and the right hand side by $\sqrt{\frac{\lambda_i}{\lambda_2}}$, then the direction of the inequality should not not change. Therefore, not only for region 2, but also for any region $i \geq 2$, we will have:

$$N \times \frac{\lambda_i}{\Sigma_{j\in\{1,...,I_0\}}\lambda_j} < \sqrt{\frac{t}{\gamma}\lambda_i}$$

Now, if we sum over all $i = 2, ..., I_0$ on both sides of the inequality above, we get:

$$N \times \Sigma_{i=2,...,I_0} \frac{\lambda_i}{\Sigma_{j\in\{1,...,I_0\}}\lambda_j} < \Sigma_{i=2,...,I_0} \sqrt{\frac{t}{\gamma}\lambda_i}$$

Rearranging, we get:

$$N \times (1 - \frac{\lambda_1}{\Sigma_{j\in\{1,...,I_0\}}\lambda_j}) < \Sigma_{i=2,...,I_0}\sqrt{\frac{t}{\gamma}\lambda_i}$$

But $N \times (1 - \frac{\lambda_1}{\Sigma_{j\in\{1,...,I_0\}}\lambda_j})$ is the total number of drivers in regions 2 through I_0. That is: $N \times (1 - \frac{\lambda_1}{\Sigma_{j\in\{1,...,I_0\}}\lambda_j}) = \Sigma_{j=2,...,I_0}\tilde{n}_j$. Therefore, we get:

$$\Sigma_{j=2,...,I_0}\tilde{n}_j < \Sigma_{i=2,...,I_0}\sqrt{\frac{t}{\gamma}\lambda_i}$$

which implies there should be at least one $j \geq 2$ such that $\tilde{n}_j < \sqrt{\frac{t}{\gamma}\lambda_j}$. But this means that for that region j, the wait time function is decreasing at $n = \tilde{n}_j$ contradicting the result that $\tilde{n}_j$ is part of an all-regions equilibrium. This completes the proof of the second observation, and hence that of lemma (A5).□

Next, we use lemma (A5) to construct an all-regions equilibrium under primitives $(\lambda, N, \frac{t}{\gamma})$. This will be a constructive proof to the existence portion of statement 4. To this end, we start from the first partial equilibrium $\bar{n}$, gradually shifting drivers from region 1 to other regions until we are left with $\tilde{n}$ drivers in region 1. That is, for any $\hat{n}_1 \in [\tilde{n}_1, \bar{n}_1]$ we consider the partial equilibrium $\hat{n} = (\hat{n}_1, ..., \hat{n}_{I_0})$ such that the $(\hat{n}_2, ..., \hat{n}_{I_0})$ is the all-regions equilibrium allocation under primitives $(\lambda_2, ..., \lambda_{I_0}), N - \hat{n}_1, \frac{t}{\gamma})$. The argument for why such partial equilibrium exists for any $\hat{n}_1 < \bar{n}_1$ is similar the argument given in proof of lemma (A4) for $\tilde{n}_1$.

Now, note that by lemma (A5), the total wait time in region 1 is larger than that in other regions when $\hat{n}_1 = \bar{n}_1$ and it is smaller in region 1 than it is in other regions when $\hat{n}_1 = \tilde{n}_1$. Therefore, there should be some $\hat{n}_1 \in [\tilde{n}_1, \bar{n}_1]$ for which the total wait time in region 1 is equal to the total wait time in all of the other regions, which themselves are equal to each other by $\hat{n}$ being a partial equilibrium the way defined above.[40] We claim such allocation $\hat{n}$ is the all-regions equilibrium of the whole market (that is, under primitives $(\lambda, N, \frac{t}{\gamma})$). The proof for this claim is as follows:

We know that under allocation $\hat{n}$ all regions have the same total wait time. We also know, by $(\hat{n}_2, ..., \hat{n}_{I_0})$ being the all-regions equilibrium under primitives $(\lambda_2, ..., \lambda_{I_0}), N - \hat{n}_1, \frac{t}{\gamma})$, that the total wait time in each region $i > 1$ is increasing at $n = \hat{n}_i$. Thus, the only thing that remains to be shown is that for $i = 1$ too the total wait time curve is increasing at $n = \hat{n_1}$. To this end, as argued before in a similar case, we need to show that $\hat{n}_1 \geq \sqrt{\frac{t}{\gamma}\lambda_1}$. Note that given $\hat{n}_1 \geq \tilde{n}_1$, it would suffice to show $\tilde{n}_1 \geq \sqrt{\frac{t}{\gamma}\lambda_1}$. We show this latter inequality by borrowing from what we already did in the proof of the last observation we made as part of proof of lemma (A5). There, we proved inequality (24) holds. Now, given that we have been assuming (without loss of generality)

[40]Note that in order to make this argument we also need to know that as we move $\hat{n}_1$ within $[\tilde{n}_1, \bar{n}_1]$, the total wait time in region 1 as well as the common total wait time in the other regions both move continuously. This is true by construction for region 1, since the total wait time function is continuous. For other regions, this needs to be shown that as we add drivers to the collection of these regions, the equilibrium total wait time moves continuously. We skip the proof of this claim here, but can provide it upon request.

that $\lambda_1 \geq \lambda_2$, and given that all λ_i are positive numbers, we get: $\frac{\lambda_1}{\lambda_2} \geq \sqrt{\frac{\lambda_1}{\lambda_2}}$. Therefore, if we multiply the left-hand side of equation (24) by $\frac{\lambda_1}{\lambda_2}$ and the right hand side by $\sqrt{\frac{\lambda_1}{\lambda_2}}$, the sign of the inequality should not change. This operation gets us:

$$N \times \frac{\lambda_1}{\Sigma_{i\in\{1,...,I_0\}}\lambda_i} \geq \sqrt{\frac{t}{\gamma}\lambda_1}$$

which is exactly what we were after. This shows that $\hat{n}$ is the all-regions equilibrium, completing the proof of the first part of statement 4 in the theorem.□

Now that we have shown the all-region equilibrium $\hat{n}$ under primitives $(\lambda, N, \frac{t}{\gamma})$ exists, we show that it indeed shows less geographical supply inequity than the old equilibrium n^*. As the first step towards this goal, note that for any $j > i > 1$, we can show the result holds based on our induction assumption. More precisely, we know that $(n_2^*, ..., n_{I_0}^*)$ is the all-regions equilibrium under primitives $((\lambda_2, ..., \lambda_{I_0}), N - n_1^*, t)$. We also know that $(\hat{n_2}, ..., \hat{n}_{I_0})$ is the all regions equilibrium under primitives $((\lambda_2, ..., \lambda_{I_0}), N - \hat{n}_1, \frac{t}{\gamma})$. The move from primitives $((\lambda_2, ..., \lambda_{I_0}), N - n_1^*, t)$ to primitives $((\lambda_2, ..., \lambda_{I_0}), N - \hat{n}_1, \frac{t}{\gamma})$ involves two steps. The first step is to divide t by some $\gamma > 1$. The second step is to add $n_1^* - \hat{n}_1$ drivers. Based on our induction assumption, both statements 4 and 5 of Theorem (1) hold for $I_0 - 1$ regions. Therefore, for any $j > i > 1$ we have:

$$\frac{\frac{\hat{n}_i}{\lambda_i}}{\frac{\hat{n}_j}{\lambda_j}} \leq \frac{\frac{n_i^*}{\lambda_i}}{\frac{n_j^*}{\lambda_j}}$$

with the inequality strict if $\lambda_i > \lambda_j$. Now the only thing that remains to show is that we can say the same not only for $j > i > 1$, but also for $j > i = 1$. In order to show this, we consider three cases.

Case 1: for every $j > 1$, we have $\hat{n}_j > n_j^*$. In this case, the result is becomes trivial given that we know $\hat{n}_1 \leq n_1^*$.

Case 2: for at least two distinct $j, j' > 1$, we have $\hat{n}_j \leq n_j^*$ and $\hat{n}_{j'} \leq n_{j'}^*$. We start with j and note that the allocation of drivers in all regions other than j –i.e., allocation $(\hat{n}_1, ..., \hat{n}_{j-1}, \hat{n}_{j+1}, ..., \hat{n}_{I_0})$ is the all-region equilibrium under primitives $((\lambda_1, ..., \lambda_{j-1}, \lambda_{j+1}, ..., \lambda_{I_0}), N - \hat{n}_j, \frac{t}{\gamma})$. Also note that allocation $(n_1^*, ..., n_{j-1}^*, n_{j+1}^*, ..., n_{I_0}^*)$ is the all-region equilibrium under primitives $((\lambda_1, ..., \lambda_{j-1}, \lambda_{j+1}, ..., \lambda_{I_0}), N - n_j^*, \frac{t}{\gamma})$. Note that the former primitives can be obtained from the latter by two moves. First, going from t to $\frac{t}{\gamma}$ for some $\gamma > 1$; and second, changing the total number of drivers from $N - n_j^*$ to the (by assumption) larger number of $N - \hat{n}_j$. Based on our induction assumptions, we know that both of these moves reduce the geographical supply inequity. Therefore, now we can claim the following for any $i \neq j$:

$$\frac{\frac{\hat{n}_1}{\lambda_1}}{\frac{\hat{n}_i}{\lambda_i}} \leq \frac{\frac{n_1^*}{\lambda_1}}{\frac{n_i^*}{\lambda_i}}$$

with the inequality strict if $\lambda_1 \neq \lambda_i$. This covers all of the comparisons that we needed with the exception of the comparison between region 1 and region j itself. But we can prove the inequality for that case as well, by going through the exact same process as above, except excluding region j' this time instead of region j. This finishes the proof of statement 4 of the theorem under case 2.

Case 3: for exactly one region $j > 1$, we have $\hat{n}_j \leq n_j^*$. In this case, we can go through the same process as that described in case 2, to show for any $i \neq j$:

$$\frac{\frac{\hat{n}_1}{\lambda_1}}{\frac{\hat{n}_i}{\lambda_i}} \leq \frac{\frac{n_1^*}{\lambda_1}}{\frac{n_i^*}{\lambda_i}}$$

with the inequality strict if $\lambda_1 \neq \lambda_i$. This time, however, we are not able to use a similar argument for the to show the result holds between regions 1 and j. The following lemmas, however, demonstrate a different way to prove the result for this specific comparison.

Lemma A6. *Under the conditions of case 3, we have $\hat{n}_1 \geq \frac{n_1^*}{\sqrt{\gamma}}$ and $\hat{n}_j \geq \frac{n_j^*}{\sqrt{\gamma}}$.*

Proof of Lemma (A6). We only show $\hat{n}_1 \geq \frac{n_1^*}{\sqrt{\gamma}}$. The argument for $\hat{n}_j \geq \frac{n_j^*}{\sqrt{\gamma}}$ is the same.

We start by observing that the total wait time w_1 in region 1 under $n_1 = \frac{n_1^*}{\sqrt{\gamma}}$ is given by:

$$\begin{aligned} w_1 &= \frac{n_1}{\lambda_1} + \frac{\frac{t}{\gamma}}{n_1} \\ &= \frac{\frac{n_1^*}{\sqrt{\gamma}}}{\lambda_1} + \frac{\frac{t}{\gamma}}{\frac{n_1^*}{\sqrt{\gamma}}} \\ &= (\frac{n_1^*}{\lambda_1} + \frac{t}{n_1^*})\frac{1}{\sqrt{\gamma}} \\ &= \frac{w^*}{\sqrt{\gamma}} \end{aligned} \tag{25}$$

where w^* is the common total wait time among all regions under primitives (λ, N, t) and the all-regions equilibrium n^* given those primitives.

Next, we show that under the *new primitives* $(\lambda, N, \frac{t}{\gamma})$, but at the *old equilibrium allocation* n^*, the total wait-time in any region i is weakly larger than $\frac{w^*}{\sqrt{\gamma}}$. To see this, we write out one such total wait time:

$$\frac{n_i^*}{\lambda_i} + \frac{\frac{t}{\gamma}}{n_i^*}$$

Note that because n^* is the all-region equilibrium under the old primitives, it must be that for all i : $\quad n_i^* \geq \sqrt{t\lambda_i}$. This gives $\frac{n_i^*}{\lambda_i} \geq \frac{t}{n_i^*}$, or, alternatively: $\frac{t}{n_i^*} \leq \frac{1}{2}(\frac{n_i^*}{\lambda_i} + \frac{t}{n_i^*}) = \frac{w^*}{2}$.

Therefore, we can write:

$$\frac{n_i^*}{\lambda_i} + \frac{\frac{t}{\gamma}}{n_i^*} = (\frac{n_i^*}{\lambda_i} + \frac{t}{n_i^*}) - \frac{t}{n_i^*}(1 - \frac{1}{\gamma})$$

$$
\begin{aligned}
&= w^* - \frac{t}{n_i^*}(1 - \frac{1}{\gamma}) \\
&\geq w^*(1 - \frac{1}{2}(1 - \frac{1}{\gamma})) \\
&= w^* \times (\frac{1 + \frac{1}{\gamma}}{2}) \\
&> w^* \times (\sqrt{1 \times \frac{1}{\gamma}}) \\
&= \frac{w^*}{\sqrt{\gamma}}
\end{aligned} \tag{26}
$$

Equations (25) and (26), together, tell us that for any i, we have $\frac{n_i^*}{\lambda_i} + \frac{\frac{t}{\gamma}}{n_i^*} > w_1$. Now notice that the total wait time under the new primitives at the old equilibrium alloction in any region is increasing. This is simply because $\forall i: \quad n_i^* \geq \sqrt{t\lambda_i} > \sqrt{\frac{t}{\gamma}\lambda_i}$. This, combined with the fact that there is at least one region i with $\hat{n}_i > n_i^*$,[41] tells us:

$$
\hat{w} \equiv \frac{\hat{n}_i}{\lambda_i} + \frac{\frac{t}{\gamma}}{\hat{n}_i} > \frac{n_i^*}{\lambda_i} + \frac{\frac{t}{\gamma}}{n_i^*} > w_1
$$

where $\hat{w}$ is defined as the common total wait time among all regions under the new primitives and new equilibrium allocation.

What $\hat{w} > w_1$ tells us is that if we reduce the number of drivers in region 1 to $n_1 = \frac{n_1^*}{\sqrt{\gamma}}$, the total wait time in region 1 falls below the equilibrium total wait time. But this means it has to be that $\hat{n}_1 > n_1 = \frac{n_1^*}{\sqrt{\gamma}}$. To see why, consider two scenarios. First, if $n_1 < \sqrt{\frac{t}{\gamma}\lambda_1}$, then by $\hat{n}_1 \geq \sqrt{\frac{t}{\gamma}\lambda_1}$, we get $\hat{n}_1 > n_1$. Next, if $n_1 \geq \sqrt{\frac{t}{\gamma}\lambda_1}$, then the wait time curve is strictly increasing when moving up from n_1, which means at some point past n_1, it hits the higher wait time $\hat{w} > w_1$. That point would be $\hat{n}_1$. Thus, the lemma has been proven for region 1. The proof for region j is exactly the same. □

We now present the another useful lemma which helps us better understand what happens to the two regions 1 and j.

Lemma A7. *Consider a market with two regions 1 and 2 only. Allocation n^* is an equilibrium in this market under primitives $(\lambda_1, \lambda_2, N, t)$ if and only if allocation $\frac{n^*}{\sqrt{\gamma}}$ is an equilibrium under primitives $(\lambda_1, \lambda_2, \frac{N}{\sqrt{\gamma}}, \frac{t}{\gamma})$.*

Proof of Lemma (A7). Follows directly from definitions.□

Now, lemmas (A6) and (A7) show us a clear way to complete the last piece of the inductive proof of statement 4 in the theorem. Based on lemma (A6), we know $\hat{n}_1 + \hat{n}_j > \frac{n_1^* + n_j^*}{\sqrt{\gamma}}$. Now define $N^* = n_1^* + n_j^*$ and $\hat{N} = \hat{n}_1 + \hat{n}_j$. We know that (n_1^*, n_j^*) was the all-regions equilibrium

[41] This is true because there are at least three regions; and besides regions 1 and j, case 3 assumes $\hat{n}_i > n_i^*$ for all i.

under primitives $(\lambda_1, \lambda_j, N^*, t)$. Thus, by lemma (A7) we can claim that $(\frac{n_1^*}{\sqrt{\gamma}}, \frac{n_j^*}{\sqrt{\gamma}})$ is the all-regions equilibrium under primitives $(\lambda_1, \lambda_j, \frac{N^*}{\sqrt{\gamma}}, \frac{t}{\gamma})$.

On the other hand, we know that $(\hat{n}_1, \hat{n}_j)$ is the all-regions equilibrium under primitives $(\lambda_1, \lambda_j, \hat{N}, \frac{t}{\gamma})$. Given that we showed $\hat{N} > \frac{N^*}{\sqrt{\gamma}}$, and given that by our strong induction assumption statement 5 is correct for all two-region cases, we can write:

$$\frac{\frac{\hat{n}_1}{\lambda_1}}{\frac{\hat{n}_j}{\lambda_j}} \geq \frac{\frac{\frac{n_1^*}{\sqrt{\gamma}}}{\lambda_1}}{\frac{\frac{n_j^*}{\sqrt{\gamma}}}{\lambda_j}} = \frac{\frac{n_1^*}{\lambda_1}}{\frac{n_j^*}{\lambda_j}}$$

with the inequality strict whenever $\lambda_1 > \lambda_j$. This completes the proof of case 3, and hence finishes the inductive proof of statement 4 of Theorem (1) with the exception of the last claim about $\gamma \to \infty$, which we turn to next.

To see why for any $i < j$ we have $\frac{\frac{n_i^{*\prime}}{\lambda_i}}{\frac{n_j^{*\prime}}{\lambda_j}} \to 1$ as $\gamma \to \infty$, assume first on the contrary, that this claim is not true. That is $\exists\, i < j$ such that $\frac{\frac{n_i^{*\prime}}{\lambda_i}}{\frac{n_j^{*\prime}}{\lambda_j}}$ does not approach 1 as $\gamma \to \infty$. We use this assumption to get a contradiction. Note that given the other claims in statement 4 of the theorem, we know that $\frac{\frac{n_i^{*\prime}}{\lambda_i}}{\frac{n_j^{*\prime}}{\lambda_j}}$ monotonically decreases as γ increases. Therefore, the only possibility for it to not approach 1, is for it to approach a number strictly above one. Denote that number by $\kappa > 1$.

Also note that as $\gamma \to \infty$, the equilibrium number of drivers in none of the regions tends to zero. This is because (i) as immediately implied by the other claims in statement 4, the number of drivers in the lowest demand region $n_I^{*\prime}$ is increasing in γ; and (ii) the number of drivers in any other region is always weakly larger than $n_I^{*\prime}$. This, along with the fact $\gamma \to \infty$ is equivalent to $t \to 0$, means that the total wait time in each region k will tend to the idle time in that region. Therefore, $\frac{\frac{n_i^{*\prime}}{\lambda_i}}{\frac{n_j^{*\prime}}{\lambda_j}} \to \kappa > 1$ implies regions i and j have different limiting total wait times at the equilibrium, which contradicts statement 1. This finishes the proof of the statement. □

Proof of Statement 6. The steps of this proof closely (almost exactly) follow the steps of the proof of statement 4. We skip it but can provide the detailed proof upon request.

Proof of Statement 5. Similar to the corresponding two-region case (i.e., proof of Proposition (5). This statement can be proven in a straightforward manner once we have proven statements 4 and 6. To be more precise, if we know that geographical supply inequity decreases in the sense defined in the statement of the theorem both (i) when we proportionally scale-up N and the vector λ and (ii) when we scale down the vector λ, it follows that the geographical supply inequity also decreases when we only scale up N, which is a certain combination of (i) and (ii).□

The above proofs show that (i) the theorem holds for $I = 2$ and that (ii) the theorem holds for any $I_0 > 2$ if it holds for all $I \in \{2, ..., I_0 - 1\}$. This means our proof is complete.■

Proof of Proposition 6. Proof of statement 1 is straightforward and is, hence, left to the reader.

Proof of Statement 2. Under primitives (λ, N, t), denote the average idle time for drivers for allocation $n \in \mathcal{N}$ by $W^{idle}(n)$. From eq. (9), we should have:

$$W^{idle}(n) \equiv \frac{\Sigma_{i=1,...,I} n_i \times \frac{n_i}{\lambda_i}}{N} \equiv \frac{1}{N} \times \Sigma_{i=1,...,I} \frac{n_i^2}{\lambda_i} \tag{27}$$

Our objective is to find the allocation n that minimizes the average wait time, subject to the restriction that $\Sigma_{i=1,...,I} n_i = N$. The first order condition requires that the gradient of $W^{idle}(n)$ with respect to n be proportional to the gradient of $\Sigma_{i=1,...,I} n_i$ with respect to n. Now note that:

$$\nabla_n(W^{idle}) \equiv \frac{2}{N} \times (\frac{n_1}{\lambda_1}, ..., \frac{n_I}{\lambda_I})$$

Also:

$$\nabla_n(\Sigma_{i=1,...,I} n_i) \equiv (1, ..., 1)$$

These two vectors being proportional implies there should be some $\xi \in \mathbb{R}$ such that:

$$\forall i \in \{1, ..., I\} : \frac{2}{N} \frac{n_i}{\lambda_i} = \xi \times 1 \tag{28}$$

This, in turn, implies:

$$\forall i, j \in \{1, ..., I\} : \frac{n_i}{\lambda_i} = \frac{n_j}{\lambda_j} \tag{29}$$

Note that there is exacly one allocation in $\mathcal{N}$ that satisfies eq. (29). That allocation is n^0. This completes the proof of statement 2 of the proposition.[42]□

Proof of Statement 3. The objective is to show $W^{idle}(n^2) \geq W^{idle}(n^1)$. Our strategy is as follows: we start from n^1 and we modify the allocation in multiple steps until we reach n^2. We show that along the way, the average idle time W^{idle} weakly increases with each one of our modifications. We first start by constructing allocation $n^{1,I}$ as follows:

- $n_I^{1,I} = n_I^2$
- $\forall i < I : n_i^{1,I} = n_i^1 \times \frac{N - n_I^2}{N - n_I^1}$

In words, allocation $n^{1,I}$ is constructed from allocation n^1 by (i) removing $n_I^1 - n_I^2$ drivers from region I,[43] and then allocating them over the other regions in a proportional manner to those regions' drivers under n^1.

[42] We skip checking the second order condition. It is fairly straightforward and is left to the reader.

[43] Note that it is straightforward to verify $n_I^1 \geq n_I^2$. This follows from the assumptions that $\forall i < j : \kappa_{ji}^2 \geq \kappa_{ji}^1$ and that $\Sigma_{i=1,...I} n_i^1 = \Sigma_{i=1,...I} n_i^2 = N$.

Remark 1. With the exception of region I, the new allocation $n^{1,I}$ preserves all of the supply ratios of n^1 for all pairs of regions. $\forall i,j < I : \frac{n_i^{1,I}}{n_i^{1,I}} = \frac{n_i^1}{n_i^1}$.

Proof of this remark is immediate. We now prove a useful lemma.

Lemma A8. $W^{idle}(n^{1,I}) \geq W^{idle}(n^1)$.

Proof of Lemma A8. Define allocation $n(z)$ by:

- $n_I(z) \equiv zn_I^2 + (1-z)n_I^1$
- $\forall i < I : n_i^{1,I} = n_i^1 \times \frac{N - n_I(z)}{N - n_I^1}$

It is straightforward to see that $n(0) = n^1$ and $n(1) = n^2$. Therefore, the lemma will be proven if we show that for all $z \in [0,1]$, the average driver idle time is weakly increasing as we increase z. To this end, we calculate $\frac{\partial W^{idle}(n(z))}{\partial z}$ and show it is non-negative for all $z \in [0,1]$. By the chain rule, $\frac{\partial W^{idle}(n(z))}{\partial z}$ is given by the following inner product:

$$\frac{\partial W^{idle}(n(z))}{\partial z} \equiv \nabla_n W^{idle}(n(z)) \cdot \frac{\partial n(z)}{\partial z} \tag{30}$$

We know:

$$\nabla_n W^{idle}(n(z)) \equiv \frac{2}{N}(\frac{n_i(z)}{\lambda_i})_{i=1,\ldots,I}$$

Also, by the definition of $n(z)$, we know:

- $\frac{\partial n_I(z)}{\partial z} \equiv -(n_I^1 - n_I^2)$
- $\forall i < I : \frac{\partial n_I(z)}{\partial z} \equiv (n_I^1 - n_I^2)\frac{n_i^1}{N - n_I^1}$

Replacing from the above formulas into eq. (30), we get:

$$\frac{\partial W^{idle}(n(z))}{\partial z} \equiv \frac{2(n_I^1 - n_I^2)}{N}[-\frac{n_I(z)}{\lambda_I} + (\Sigma_{i=1,\ldots,I-1}\frac{n_i^1}{N - n_I^1}\frac{n_i(z)}{\lambda_i})] \tag{31}$$

To show the above is non-negative, we claim that:

$$\forall i < I : \frac{n_i(z)}{\lambda_i} \geq \frac{n_I(z)}{\lambda_I} \tag{32}$$

To see why eq. (32) holds, note that for all $z \in [0,1]$, we have $n_I(z) \leq n_I^1$ and $\forall i < I : n_i(z) \geq n_i^1$. Therefore, we can write:

$$\frac{n_i(z)}{n_I(z)} \geq \frac{n_i^1}{n_I^1}$$

But we know that by assumption:

$$\frac{n_i^1}{n_I^1} \geq \frac{\lambda_i}{\lambda_I}$$

Combining the above two inequalities, we get $\frac{n_i(z)}{n_I(z)} \geq \frac{\lambda_i}{\lambda_I}$. Re-arranging, we get eq. (32).

Next, replacing from eq. (32) into eq. (31) and noting that $n_I^1 - n_I^2$ is non-negative, we have the following for any $z \in [0,1]$:

$$\frac{\partial W^{idle}(n(z))}{\partial z} \geq \frac{2(n_I^1 - n_I^2)}{N}[-\frac{n_I(z)}{\lambda_I} + (\Sigma_{i=1,\ldots,I-1}\frac{n_i^1}{N - n_I^1}\frac{n_I(z)}{\lambda_I})] \tag{33}$$

$$= \frac{2(n_I^1 - n_I^2)}{N}\frac{n_I(z)}{\lambda_I}[-1 + (\Sigma_{i=1,\ldots,I-1}\frac{n_i^1}{N - n_I^1})]$$

$$= \frac{2(n_I^1 - n_I^2)}{N}\frac{n_I(z)}{\lambda_I}[-1 + (\frac{\Sigma_{i=1,\ldots,I-1} n_i^1}{N - n_I^1})]$$

$$= \frac{2(n_I^1 - n_I^2)}{N}\frac{n_I(z)}{\lambda_I}[-1 + (\frac{N - n_I^1}{N - n_I^1})]$$

$$= \frac{2(n_I^1 - n_I^2)}{N}\frac{n_I(z)}{\lambda_I}[0]$$

$$= 0$$

Therefore, it follows that $W^{idle}(n^{1,I}) \geq W^{idle}(n^1)$, completing the proof of Lemma A8.□

Next, we start from $n^{1,I}$ and make another adjustment in order to get to n^2. Informally, we take $n^{1,I}$, move $n_{I-1}^{1,I} - n_{I-1}^2$ drivers from region $I-1$, and proportionally re-allocate them to regions 1 through $I-2$.[44] Note that we are not modifying anything about region I any more. Therefore, the new allocation will have exactly as many drivers as n^2 does in regions $I-1$ and I; but in other regions, its drivers will be allocated proportionally to what n^1 and, by Remark 1, $n^{1,I}$ had.

We denote this new allocation $n^{1,I-1}$. Before formally defining it, we introduce another notation. Denote $N^I = \Sigma_{i=1,\ldots,I-1} n_i^{1,I} = N - n_I^2$. Now we formally define the allocation $n^{1,I-1}$ and how it is constructed from $n^{1,I}$ as follows:

- $n_I^{1,I-1} = n_I^{1,I} = n_I^2$
- $n_{I-1}^{1,I-1} = n_{I-1}^2$
- $\forall i < I-1 : n_i^{1,I-1} = n_i^{1,I} \times \frac{N^I - n_{I-1}^2}{N^I - n_{I-1}^{1,I}}$

Next, in a similar manner to Lemma A8, we can show $W^{idle}(n^{1,I-1}) \geq W^{idle}(n^{1,I})$. We skip the steps here because they are exactly the same as the steps of the proof in Lemma A8. We then continue constructing allocations $n^{1,I-2}, n^{1,I-3}$ etc. in a similar manner, and in every step m, we

[44] Again, note that $n_{I-1}^{1,I} - n_{I-1}^2 \geq 0$. To see why this is true, observe that (i) $\Sigma_{i=1,\ldots,I-1} n_i^{1,I} = \Sigma_{i=1,\ldots,I-1} n_i^2$, and (ii) by combining Remark 1 with the assumptions in the proposition: $\forall i < j < I : \frac{n_i^{1,I}}{n_j^{1,I}} \geq \frac{n_i^2}{n_j^2}$.

show $W^{idle}(n^{1,I-m}) \geq W^{idle}(n^{1,I-m+1})$. Given that our construction of such allocations implies $n^{1,1} \equiv n^2$, then we get:

$$W^{idle}(n^2) = W^{idle}(n^{1,1}) \geq W^{idle}(n^{1,2}) \geq ... \geq W^{idle}(n^{1,I}) \geq W^{idle}(n^1)$$

which finishes the proof of Statement 3, and that of the proposition.■