Addressing the minimum fleet problem in on-demand urban mobility

M. M. Vazifeh,[1] P. Santi,[2] G. Resta,[3] S. H. Strogatz,[4] and C. Ratti[1]

[1]*Senseable City Laboratory, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA*

[2]*Senseable City Laboratory, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA – Istituto di Informatica e Telematica del CNR, Via G. Moruzzi 1, 56124 Pisa, Italy*

[3]*Istituto di Informatica e Telematica del CNR, Via G. Moruzzi 1, 56124 Pisa, Italy*

[4]*Dept of Mathematics, Cornell University, 533 Malott Hall, Ithaca, NY 14853, USA*

Information and communication technologies have opened the way to new solutions for urban mobility that provide better ways to match individuals with on-demand vehicles. However, a fundamental unsolved problem is how to best size and operate a fleet of vehicles, given a certain demand for personal mobility. Previous studies [1–5] either fail to provide a scalable solution or require changes in human attitudes towards mobility. Here, we provide a network-based solution to the following 'minimum fleet problem': given a collection of trips (specified by origin, destination, and start time), what is the minimum number of vehicles needed to serve all the trips without incurring any delay to the passengers? By introducing the notion of a 'vehicle sharing network', we present an optimal computationally efficient solution to the problem, as well as a nearly optimal solution amenable to real-time implementation. We test both solutions on a data set of 150 million taxi trips taken in the city of New York over one year [6]. The real-time implementation of the method with near-optimal service levels allows a 30% reduction in fleet size compared to current taxi operation. These improvements follow from a mere re-organization of taxi dispatching that could be implemented with a simple urban app; they do not assume ride sharing [7–9], nor require changes to regulations, business models, or human attitudes toward mobility to become effective. Our results could become even more relevant in the years ahead as fleets of networked, self-driving cars become commonplace [10–14].

Two trends – the rise of the autonomous and connected car, and the emergence of a 'sharing economy' [10, 11] of transportation – seem poised to revolutionize the way personal mobility needs will be addressed in cities. The way current modes of transportation such as the private car, taxi, or bus operate will be challenged and increasingly replaced by personalized, on-demand mobility systems operated by vehicle fleets, similar to what companies like UBER and Lyft have started to offer. If such trends continue, they could lead to the displacement, or eventual disappearance, of jobs for bus and taxi drivers. Along with these possible societal costs, the transportation revolution could also offer immense benefits, including opportunities to resolve existing inefficiencies in individual urban mobility [12–14], thereby reducing traffic, whose carbon footprint currently accounts for about 23% of global greenhouse gas emissions [15, 16].

To turn these opportunities into tangible environmental and societal benefits, autonomous and on-demand mobility systems need to be designed and optimized for efficiency, and integrated with carbon-efficient public transport. This requires the definition of models and algorithms for the evaluation of shared mobility systems that are both *computationally efficient* and *accurate*. The former property is mandated by the need to cope with hundreds of thousands (or sometimes millions) of trips routinely occurring in a large city. The latter property determines the relevance of the model results to the real world.

In what follows, we solve the 'minimum fleet problem' for the general case of on-demand mobility, and show that its solution for a specific case – taxi trips – could lead to breakthroughs in operational efficiency as compared to current operation.

To the best of our knowledge, no publicly available solution currently exists to address this minimum fleet-size problem at the urban scale for on-demand mobility in both private and public sectors. On one hand, accurate methods based on mathematical programming (as traditionally used in the design of transportation systems [1–5, 9]) can handle only a few thousand trips or vehicles at most, which is well below the hundreds of thousands or even millions of trips or vehicles routinely operating in large cities. On the other hand, city-scale studies like the one reported in [17] are obtained using a model of transportation based on aggregated mobility data and Euclidean spatial assumptions, and hence lack the resolution necessary to accurately estimate the urban-scale benefits of vehicle sharing.

The study reported herein starts from the notion of the shareability network introduced in [7], which did not focus on the dispatching of vehicles. The type of shareability network introduced here is new and profoundly different from that studied previously: it models the sharing of *vehicles*, whereas previous networks studied by [7–9] modeled the sharing of *rides*. The main methodological contribution of this letter is showing how this vehicle sharing network can be translated into an *exact* formulation of the minimum fleet problem as a minimum path cover problem on directed graphs, thus establishing a connection to the rich applied mathematics and computer science field of graph algorithms. Along with a structural property of vehicle sharing networks herein proved, this connection allows the derivation of computationally efficient

algorithms for optimal vehicle deployment and dispatching. While optimally solving the minimum fleet size problem requires offline knowledge of daily mobility demand, in the following we also present a near-optimal, online version of the algorithm that can be executed in real time knowing only a small amount of the trip demand.

We are given a collection $\mathcal{T}$ of individual trips representing a portion of urban mobility demand during a certain time interval, such as a day. Each trip $T_i \in \mathcal{T}$ is defined as a tuple $(t_i^p, t_i^d, l_i^p, l_i^d)$ where t_i^p represents the desired pick-up time, l_i^p the pick-up location, t_i^d the drop-off time, and l_i^d the drop-off location, respectively. Here, the pick-up time means the earliest time t_i^p at which the passenger can be picked up at location l_i^p. The drop-off time means the *estimated* time of dropping off the passenger where the estimate is done using a travel time estimation model and assuming the passenger leaves the pick-up location at time t_i^p. In contrast to [17], travel times here are computed using the actual road network, and using GPS-based estimations derived from the taxi trip data set that account for hourly variations in traffic as done in [7]. In case the set $\mathcal{T}$ is extracted from a real world data set (e.g., taxi trips), the times t_i^p and t_i^d represent the actual time at which a passenger is picked up and dropped off, respectively.

The minimum fleet problem is formally defined as follows: "Find the minimum number of vehicles needed to serve all trips in $\mathcal{T}$, given that a vehicle is available at each l_i^p on or before t_i^p." A service designed around this problem is ideal from a passenger's perspective, since a vehicle is guaranteed to be available at the desired location and time. On the other hand, the above problem formulation might entail significant inefficiencies for the operator and the environment. Consider two consecutive trips T_A, T_B served by a single vehicle, and call the time needed to connect them the *(trip) connection time*, formally $t_{AB} = t_B^p - t_A^d$. If this time is very long, say, a few hours, it is trivially possible to connect trips that occur at distant locations or times. Hence, an excessively large connection time leads to inefficiencies for the operator (longer traveled distances, lower vehicle occupancy ratio) and the citizens (a lot of emissions and traffic just to connect trips). We therefore re-formulate the problem as follows: "Find the minimum number of vehicles needed to serve all trips in $\mathcal{T}$, under the assumptions that $a1$) a vehicle is available at each l_i^p on or before t_i^p and $a2$) the connection time is at most δ minutes", where the upper bound δ on the connection time is a problem parameter.

Fig. **1** illustrates the construction of the *vehicle shareability network*, that enables optimally solving the minimum fleet problem with parameter δ. This is a *directed* network defined as $V = (N, E)$, where node $n_i \in N$ corresponds to trip $T_i \in \mathcal{T}$ and the directed edge $(n_i, n_j) \in E$ if and only if $(t_i^d + t_{ij}) \leq t_j^p$ (which accounts for assumption $a1$)) and $t_j^p - t_i^d \leq \delta$ (which accounts for assumption $a2$)). Here, t_{ij} represents the estimated travel time between l_i^d and l_j^p. The existence of a link in the network indicates that the two incident trips can be consecutively served by a single vehicle, and a path in V corresponds to a sequence of trips that can be served by a single vehicle – a *dispatching*. Therefore, solving the minimum fleet problem is equivalent to finding the number of paths (vehicles) in the minimum path cover of V. The solution also gives the optimal dispatching strategy, i.e., a sequence of trips to be served for each vehicle in the minimum fleet. The problem of finding the minimum path cover on general graphs is NP-hard, but it can be solved efficiently on directed acyclic graphs [18] using the Hopcroft-Karp algorithm for bi-partite matching [19]. The acyclic nature of time guarantees that any vehicle shareability network is a directed acyclic graph, and the minimum fleet problem can be efficiently and optimally solved – see Methods section for formal proofs.

We have tested our methodology on a data set of over 150 million taxi trips performed in the city of New York in the year 2011. This data set has been selected among a number of available data sets [8] because it is publicly available and, thanks to taxi statistics published by the New York Taxi and Limousine Commission [6], it is possible to directly compare our methodology with current taxi operation [20]. The data has been sliced into daily data sets $\mathcal{T}_i$, each of which is an input to the minimum fleet size problem.

Next, we discuss how to set the parameter δ. When δ is decreased to 0, we approach a situation in which each trip is served by a dedicated vehicle: a solution with maximal vehicle utilization which is optimal also for traffic – under the assumption that vehicles materialize at the origin and dematerialize at the destination of the served trip – but incurring prohibitive costs for the mobility operator. On the other hand, when δ grows excessively, fleet size is reduced, but the operational and traffic efficiency problems described previously occur. Thus, the setting of δ is an important design choice that shall be left in the hands of mobility operators, traffic authorities and policy makers. In this study, we set $\delta = 15\ min$ as explained in the Methods section. The results of our method with different values of δ are reported in the Methods section (see Extended Data Fig. **1**).

Fig. **2** shows the daily number of vehicles needed to address the entire taxi demand in New York using our approach. The minimum number of vehicles needed to serve trips is correlated with the number of daily trips (see Fig. **2a**), with an overall R^2 value of 0.74. However, for the vast majority of days having between 300,000 and 550,000 trips (Fig. **2a**, inset) this correlation becomes much weaker, with an R^2 value of only 0.18. Thus, trip density is a first determinant of fleet size, but trip spatio-temporal patterns are likely to play a strong role as well. To investigate this issue further, we have analyzed daily vehicle usage in the optimal solution.

The vehicle usage analysis reported in the Methods section shows that a fraction of vehicles ranging between 5 and 10% are highly underutilized and serve only around 1% of the trips, a lower utilization pattern that occurs especially during the weekend and is likely related to the extra nightly demand. The analysis also highlighted clear weekly patterns in vehicle utilization, consistent with the relatively stable vehicle fleet size

across the year. This observed stability can be explained by a simple model for vehicle-trip assignment, and is fundamental for mobility operators: it indicates that investment in acquiring an optimal number of vehicles for operation gives consistent yearly returns. The dip in vehicle fleet size occurring in the weekend hints also to an opportunity for performing routine vehicle maintenance on a weekly basis.

A better scaling law relating vehicle fleet size with daily number of trips can be obtained by defining a metric for fleet sizing that incorporates for how long a vehicle is used during a day. We define a 'full time equivalent' vehicle as a vehicle continuously operating $24h$ a day [21]. Fig. **2b** shows that the scaling law relating number of daily trips with full time equivalent vehicles is more accurate than the previous one, with an R^2 value increased from 0.74 to 0.91, and from 0.18 to 0.70 for the prevailing trip intense days reported in the inset.

Fig. **3** shows the efficiency breakthrough provided by network-based optimization: when compared to current taxi operation in New York, the number of circulating taxis can be reduced by an impressive 40%, and kept fairly constant through the day. This improvement is all the more noticeable considering that it is achieved without imposing any delay on customers, nor sharing of rides as in [7, 9]. The fact that fleet size can be reduced of as much as 40% *without the use of ride sharing* and with *no delay for passengers* has, to our best knowledge, never been reported in the literature before, and it is one of the main results of this paper.

The 40% fleet reduction reported above refers to the model with full knowledge of daily trip demand. If only a portion of trip demand is known, as in current on-demand mobility services where trip requests are collected in real-time, we can still achieve near-optimal performance with the on-line version of the algorithm reported in the Method section. This version collects trip requests for a short time, e.g., one minute, and locally optimizes vehicle dispatching based on this limited knowledge. Fig. **4** shows that, with a 30% fleet reduction and *online operation*, more than 90% of the trip requests can be successfully served, hitting a performance very close to the 40% fleet reduction possible when the entire daily demand is known beforehand.

Our approach assumes that trips requests and vehicle dispatching decisions are centralized, a model which is radically different from current taxi operation and similar to the one used by online mobility operators. Therefore, the benefits of optimized operation reported in Fig. **3** can be interpreted as being implied by the transition from a fully distributed operation where the deployment strategy is based on individual driver decisions, to a centralized operation where dispatching decisions are globally optimized. To some extent, our results can then be seen as a quantification of the well-known game-theoretic notion of *price of anarchy* [22] in urban taxi operation. Taking a mobility market perspective, this is a transition from a regulated mobility market with numerous micro-operators (down to the level of the single taxi driver), to a monopolistic market with a single mobility operator with centralized operation. While optimal from the vehicle operation and environmental viewpoint, a monopolistic market is however highly undesirable for many other reasons, most importantly lack of competition and consequent higher prices for customers. An additional analysis reported in the Methods section shows that most of the efficiency benefits of centralized vehicle operation are still possible in an oligopolistic market.

While the characterization of minimum fleet size reported herein is fully representative of an autonomous driving scenario where human operation of vehicles is not necessary, constraints on driver availability and maximum operating hours, shift operation, and so on, might imply relatively larger values of the minimum fleet requirement with respect to those predicted herein. While beyond the scope of this paper, extending the concept of vehicle sharing network to incorporate driver constraints is possible and is left for further analysis.

A 30% reduction in taxi fleet size as predicted by our model has the positive effect of reducing taxi-induced traffic, which already represents a large fraction of urban traffic in some cities such as New York. Broader effects on traffic are foreseen would our methodology be used for optimizing urban "on-demand" mobility services more in general, especially in a future of autonomous vehicles. On the other hand, it is well-known that the improvement in mobility efficiency are sometimes linked with an increase in demand which, in turn, could reduce the amount of traffic reductions. Evaluating this "second-order" effect of optimized fleet operation on urban traffic requires coupling a micro-level traffic simulation, agent-based passenger models, and our network-based methodology, a challenging task which is left for a future work.

We finally observe that, while applied herein to taxi trips as a case study, the proposed methodology for optimal vehicle fleet sizing and dispatching is very general and can be applied to model any type of point-to-point mobility. However, the presented approach focuses on optimizing and dispatching of a single fleet of vehicles. Optimization across different fleets and transportation modes is possible by extending our approach to consider multiple coexisting fleets of various types to serve the mobility demand. With the approaching advent of autonomous mobility, and the forecast increase in sharing cars (or other autonomous vehicles such as flying drones) the problem of how to optimize and orchestrate multiple autonomous fleets will come to the forefront, and might be addressed using the scalable and accurate analytical tools presented herein for optimally solving the "minimum fleet" problem.

Online Content: Methods, along with any additional Extended Data display items, are available in the online version of the paper; references unique to these sections appear only in the online paper.

[1] Berbeglia G, Cordeau JF, Laporte G (2010) Dynamic pick-up and Delivery Problems. *Eur. J. Op. Res.* 202(1): 8-15.

[2] Laporte G (1992) The vehicle routing problem: An overview of exact and approximate algorithms. *European journal of operational research* 59(3), 345-358.

[3] Baker BM, Ayechew MA (2003) A genetic algorithm for the vehicle routing problem. *Computers and Operations Research* 30(5), 787-800.

[4] Yang J, Jaillet P, Mahmassani H (2004) Real-time multivehicle truckload pick-up and delivery problems. *Transp. Sci.* 38(2):135–148.

[5] Clare GL, Richards AG (2011) *Optimization of taxiway routing and runway scheduling*. in *Proc. IEEE Intelligent Transportation Systems* 12(4):1000-1013.

[6] Bloomberg MR, Yassky D (2014) *New York City Factbook* Taxi and Limousine Comission

[7] Santi P, Resta G, Szell M, Sobolevsky S, Strogatz SH, Ratti C (2014) Quantifying the benefits of vehicle pooling with shareability networks. *Proceedings of the National Academy of Sciences.* 111(37):13290-13294.

[8] Tachet R, Segarra O, Santi P, Resta G, Szell M, Strogatz SH, Ratti C (2017) Scaling Law of Urban Ride Sharing. *Nature Scientific Reports.* 7: srep42868.

[9] Alonso-Mora J, Samaranayake S, Wallar A, Frazzoli E, Rus D (2017) On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment. *Proceedings of the National Academy of Sciences*, 114(3): 462-467.

[10] Rosenberg T (2013) It's not just nice to share, it's the future. *N.Y. Times: Opinionator*, https://perma.cc/89YT-VHVF

[11] Sundararajan A (2013) From Zipcar to the sharing economy. *Harvard Business Review*, https://hbr.org/2013/01/from-zipcar-to-the-sharing-eco

[12] Mitchell W.J., Borroni-Bird, C.E., Burns, L.D. (2010) Reinventing the automobile: personal urban mobility for the 21st century. *The MIT Press*, Cambridge, US.

[13] Botsman R, Rogers R (2010) *What's mine is yours: The rise of collaborative consumption* (HarperCollins).

[14] Handke V, Jonuschat H (2013) *Flexible Ridesharing* (Springer).

[15] United Nations Environment Programme. (2010) Annual report. *www.unep.org/annualreport/2010/pdfs/UNEP-AR-2010-FULL-REPORT.pdf*

[16] Barth M, Boriboonsomsin K (2011) Real-world carbon dioxide impacts of traffic congestion. (2011) *Transportation Research Record: Journal of the Transportation Research Board* 2058: 163-171.

[17] Spieser K, Treleaven K, Zhang R, Frazzoli E, Morton D, Pavone M (2014) Toward a systematic approach to the design and evaluation of automated mobility on-demand systems: a case study in Singapore. *Road Vehicle Automation* Springer International Publishing, 229-245.

[18] Boesch FT, Gimpel JF (1977) Covering the points of a digraph with point-disjoint paths and Its application to code optimization. in *Journal of the ACM* Vol. 24, n.2, pp. 192–198.

[19] Hopcroft J, Karp R (1973) An $n^{\frac{5}{2}}$ algorithm for maximum matching in bipartite graphs. in *SIAM Journal on Computing* Vol. 2, n. 4, pp. 225-231.

[20] The methodology proposed herein has been tested on a similar dataset covered by NDA, obtaining similar results.

[21] In case of human driven vehicles, we can think of having the vehicle operated on three 8 hours shifts, for instance.

[22] Roughgarden T (2005) *Selfish routing and the price of anarchy* MIT Press

Acknowledgments. P. S, M.V., and C.R. would like to gratefully thank; UBER, Allianz, the Amsterdam Institute for Advanced Metropolitan Solutions, Ericsson, the Fraunhofer Institute, Liberty Mutual Institute, Philips, the Kuwait-MIT Center for Natural Resources and the Environment, Singapore-MIT Alliance for Research and Technology (SMART), Volkswagen Electronics Research Laboratory, and all the members of the MIT Senseable City Lab Consortium for supporting this research.

Author Contributions. P. S. defined the problem, designed the solution and algorithms, and contributed to the analysis and paper writing. M.V. designed and performed the analysis, developed models and simulations and wrote the paper. G.R. contributed to the algorithm design, implemented the algorithms, and contributed to the analysis. S.H.S. contributed to writing. C.R. supervised the research and contributed to writing.

Data and code availability. All data processed during the course of this study are included in this Letter and its Supplementary Information. The code for generating the shareability networks and optimal dispatching is subject to licensing and could be made available upon request to the authors. NYC taxi data used in the study can be downloaded at `http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml`

Author information. The authors have no competing financial interests. Correspondence and requests for material should be addressed to M. M. Vazifeh.

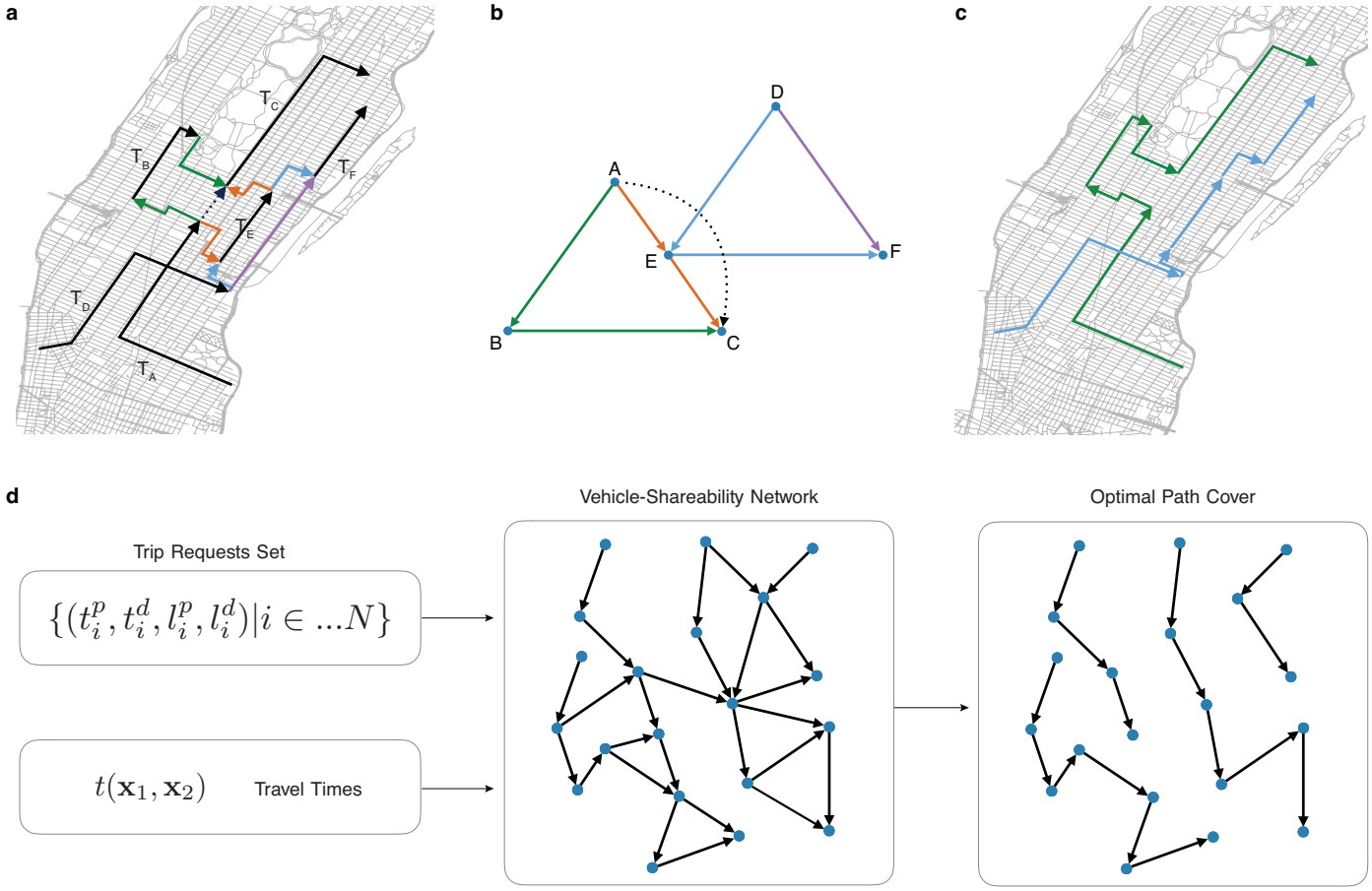


Figure 1. Constructing vehicle-shareability network.

406

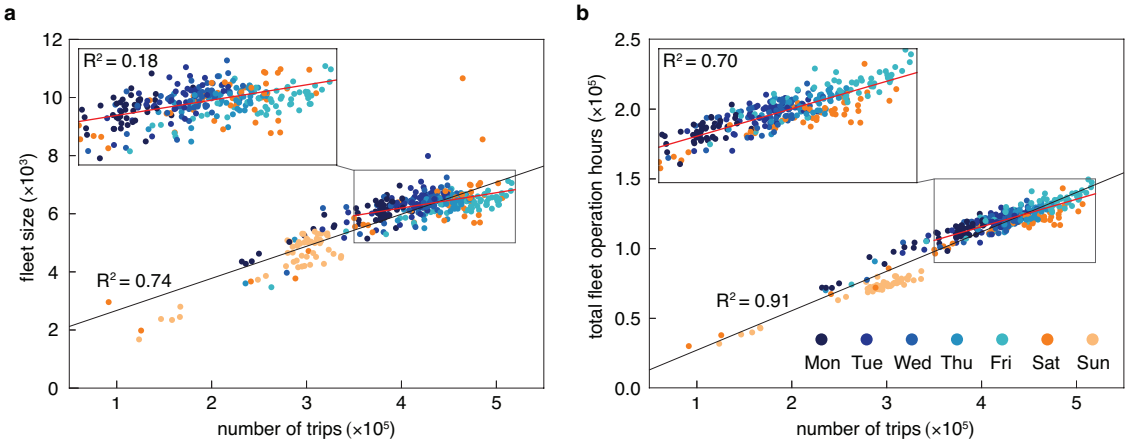


Figure 2. Minimum fleet-size analysis.

407

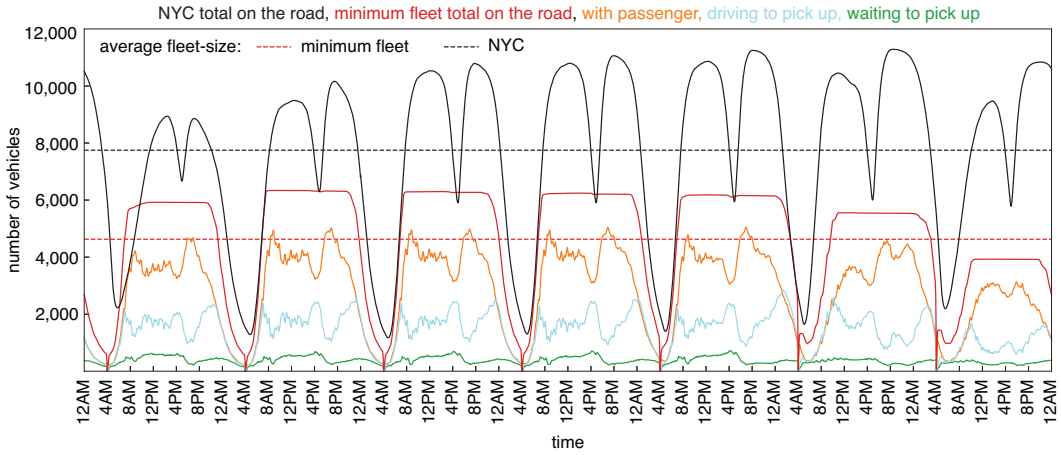


Figure 3. Fleet efficiency comparison.

408

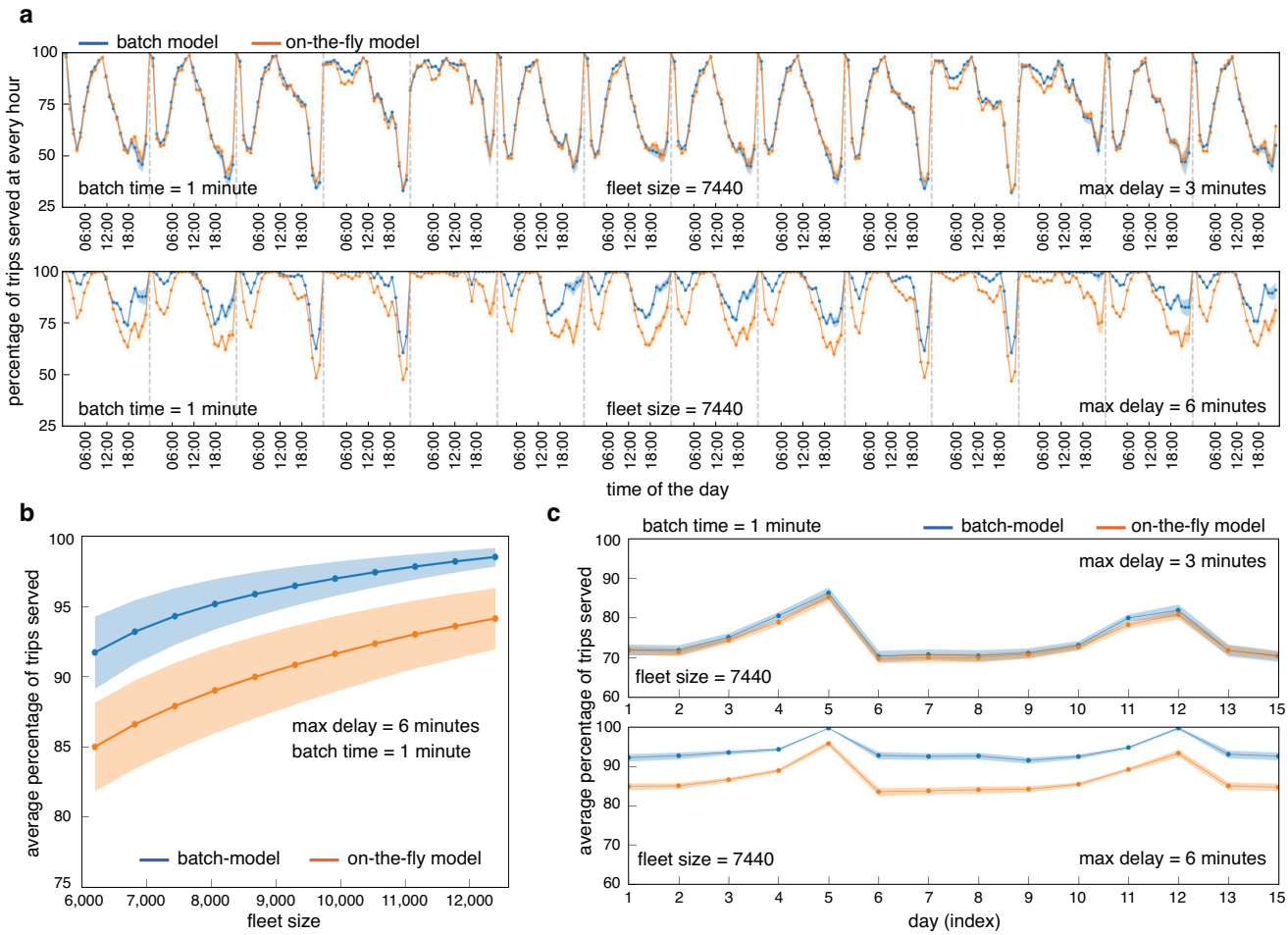
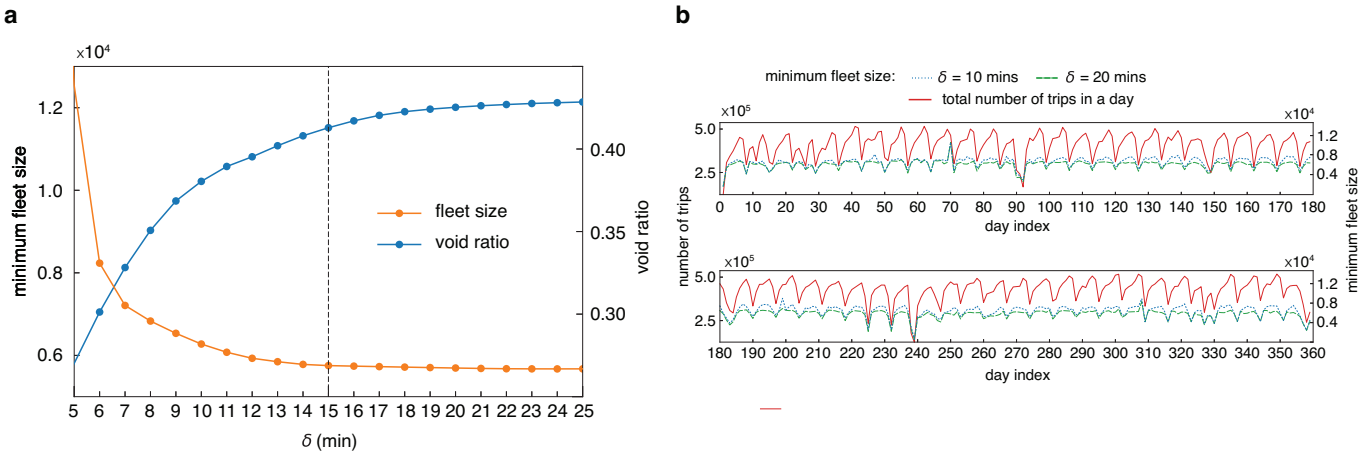


Figure 4. Performance of the network-based online vehicle dispatching model.

409

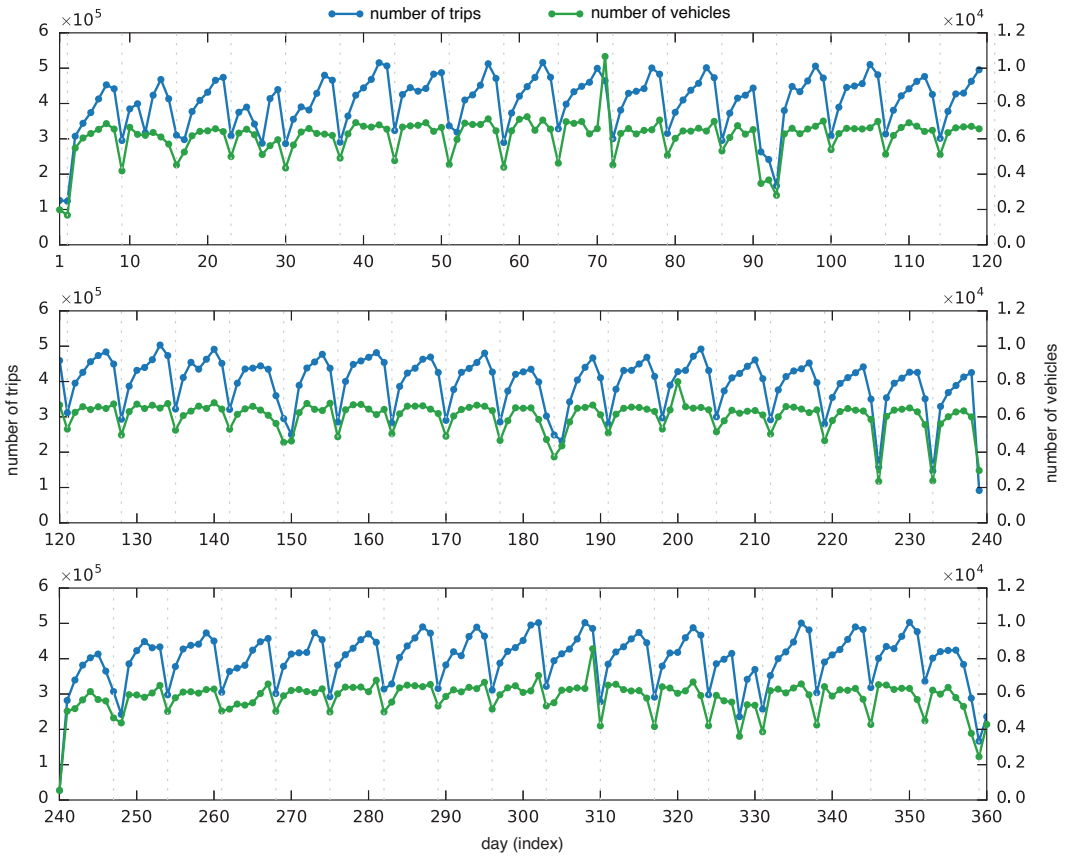
x	Δt (minutes)	$t_g(ms)$ average	$t_g(ms)$ max	$t_{tva}(ms)$ average	$t_{tva}(ms)$ max
1.2	2	12	27	2	4
	3	14	29	3	8
	4	15	30	4	16
	5	17	37	6	29
	6	20	49	8	46
2.0	2	26	42	5	7
	3	28	46	7	14
	4	33	51	12	28
	5	41	68	20	55
	6	50	93	32	100

Extended Data Table 1. **Real-time computation runtimes in milliseconds.**



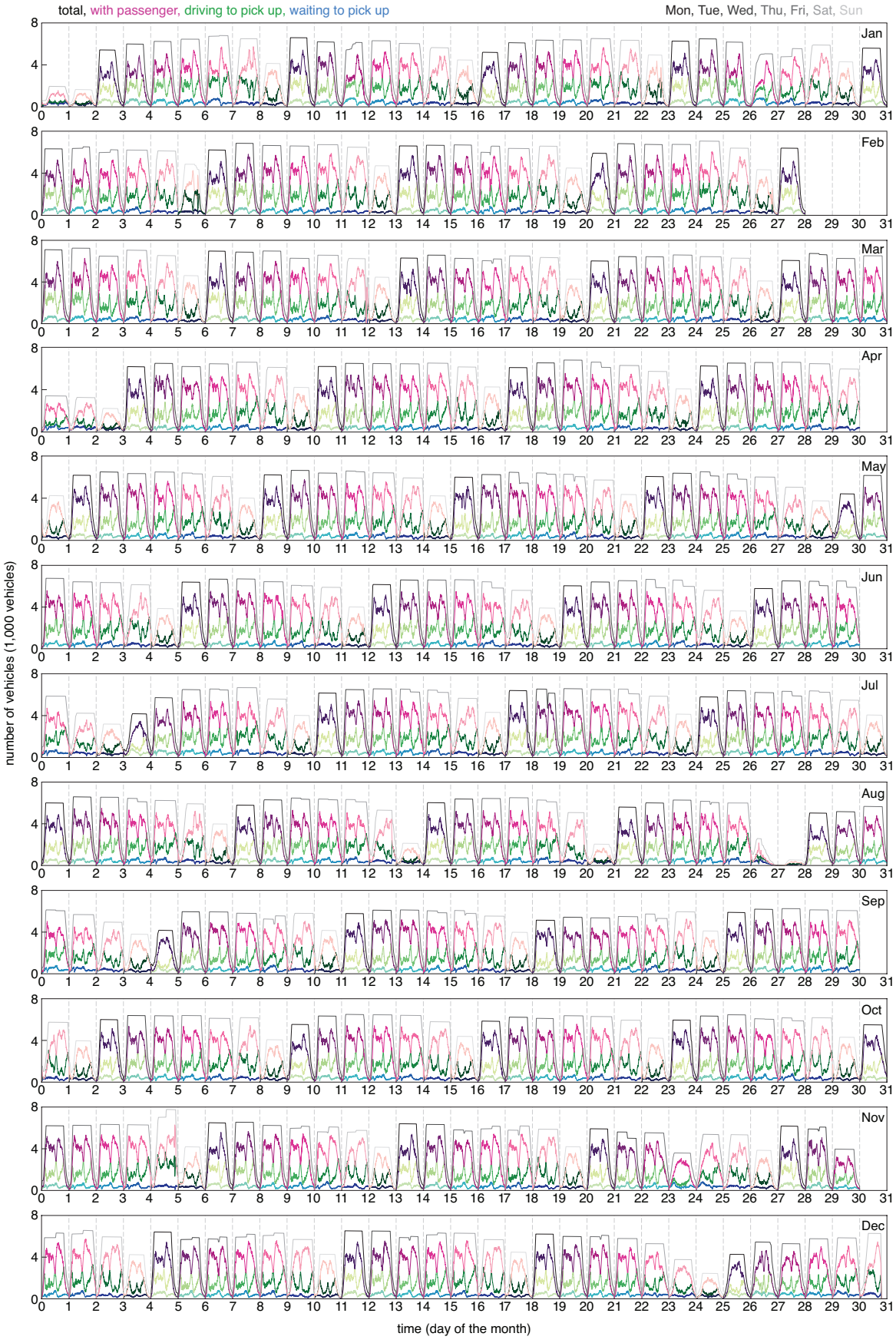
Extended Data Figure 1.

411

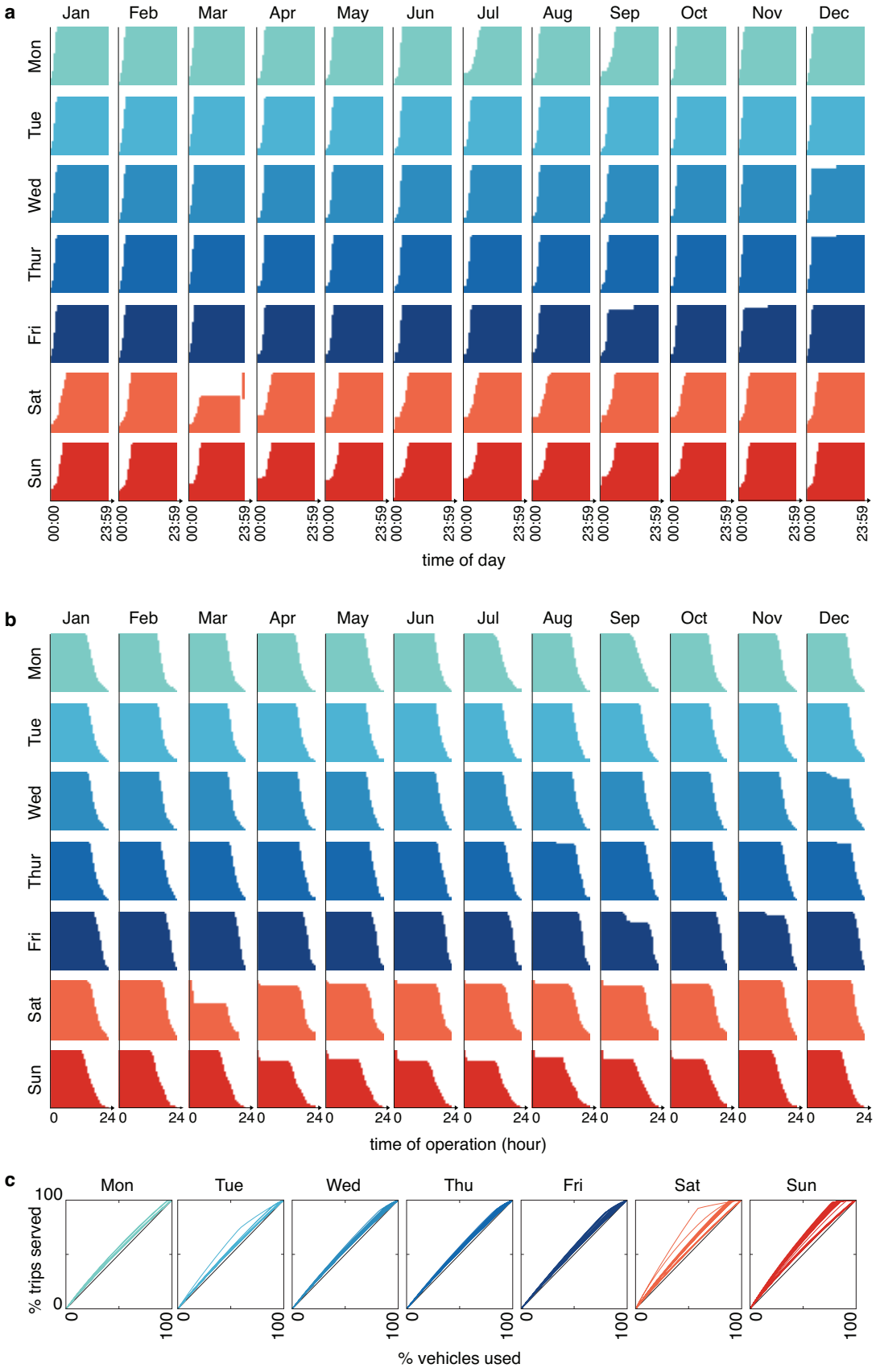


Extended Data Figure 2.

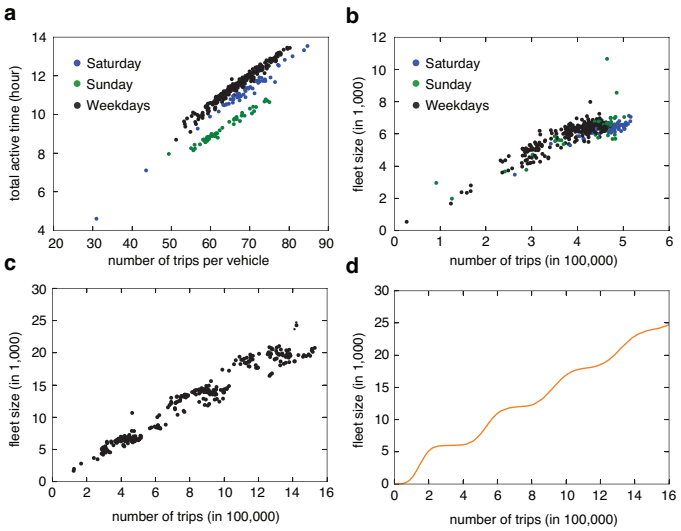
412



Extended Data Figure 3.

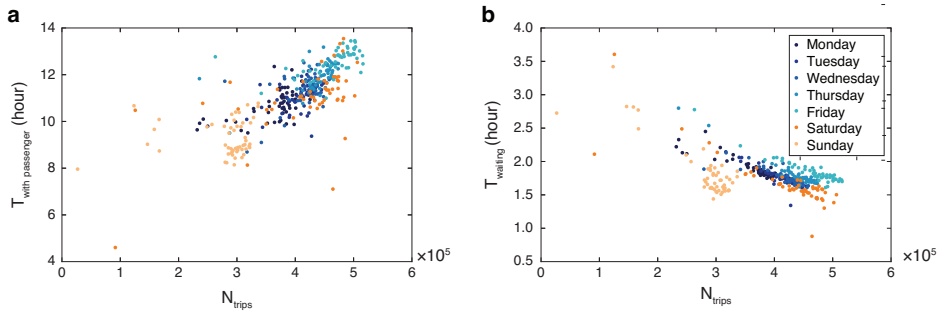


Extended Data Figure 4.



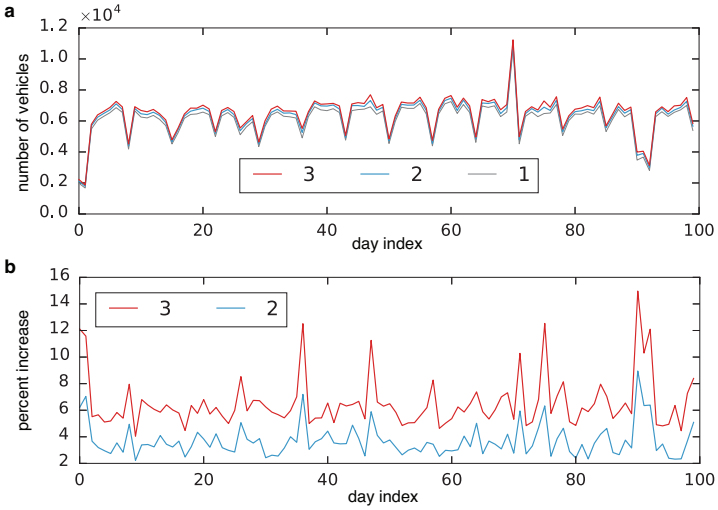
Extended Data Figure 5.

413

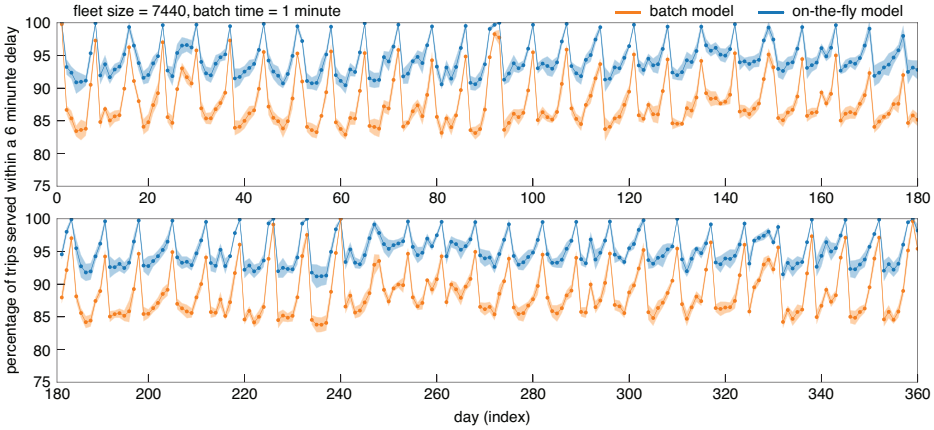


Extended Data Figure 6.

414



Extended Data Figure 7.



Extended Data Figure 8.

Figure 1. **Constructing vehicle-shareability network. a**, Several trip requests have been depicted on the map index alphabetically as T_A ... T_F. The colored path on the map represents different possibilities for vehicle dispatching. **b** Colors have been used in the vehicle-shareability network to specify how various dispatching can be represented using paths on this network. One of the two dispatching scenarios required only two vehicles while the other required three. **c** The optimal vehicle dispatching routes are represented on the map. **d** To construct the vehicle-shareability network, trip set $\mathcal{T}$ and the travel times are taken as inputs. Two trips are connected by a directed edge if enough gap in time between the drop-off of the first and pick-up point of the next trip exists to allow a single vehicle to travel between the two points before the pick-up time of the second trip starts according to the travel-time information. Furthermore, the upper bound δ on the trip connection time must not be exceeded. The path covering algorithm yields the path set that covers the entire node set, ensuring that all trips are served, while minimizing the number of paths (vehicles) in the solution. This is the optimal solution to the minimum fleet problem with parameter δ.

Figure 2. **Minimum fleet-size analysis. a**, The interplay between the number of daily trips and the minimum number of vehicles required to serve them. The color of the dots corresponds to different weekdays. Clustering of points with the same color suggest occurrence of weekly patterns, that are confirmed by yearly analysis reported in the Methods section (see Extended Data Fig. **2**). The plot shows moderate correlation between the two quantities. However, when restricting the attention to days with the number of trips ranging from 350,000 to 550,000 (inset), the correlation becomes much weaker, as the fleet size remains within a narrow window around 6,000 vehicles. **b** The correlation between the number of daily trips and the total fleet operation hours. The latter is defined as the total time of operation summing over the operation times for each vehicle in the minimum fleet obtained for a day. The operation time for each vehicle is defined as the total time a vehicle is operating on the road to serve the trips. Total Fleet operation hours manifests a much stronger correlation with the number of trips than what the number of vehicles shows.

Figure 3. **Fleet efficiency comparison.** Comparison between actual number of NYC taxis on the road [6] (black curve), versus the minimum number of vehicles as computed by our optimal approach (red curve). For the optimized approach, a breakdown of vehicles into occupied, waiting for passenger, and driving to next pick-up is also reported. The curves are computed averaging data across one year. As shown by the dotted lines corresponding to average deployment, optimized dispatching brings a reduction in number of circulating taxis from 7,748 down to 4,627, a 40% reduction. With online operation ready implementable into a smartphone app, our method brings a near-optimal 30% reduction in number of circulating taxis (see Fig. **4**).

Figure 4. **Performance of the network-based online vehicle dispatching model**. **a**, Two-panel plot comparing the percentage of trips served within a certain maximum delay $max(\Delta t)$ based on the batch-based optimized dispatching (blue line) vs sequential dispatching as described in the Methods section. The panel at the top represents the results for $max(\Delta t) = 3mins$ and the one at the bottom is for $max(\Delta t) = 6mins$. The use of network-based dispatching (blu line) provides a substantial improvement in the percentage of successfully served trips with respect to sequential dispatching (orange line). The fleet-size is set to $N = N_{min}.x$ where N_{min} is the average minimum fleet size obtained from the oracle model based on the historical data for the 15 days considered in the evaluation. The fleet size inflation ratio x is set to 1.2, the maximum passenger waiting time is set to $max(\Delta t) = 3, 6\ min$, and the batching time t_{batch} in batch-based dispatching is set to $1\ min$. **b**, The average percentage of trips successfully served within 6 minutes delay is above 90% for a fleet of 6200 vehicles. Similar performance can only be achieved in the sequential dispatching when the fleet size is more than 10,000. **c**, Plots showing daily averages of the total percentage of trips in a day successfully served within the specified delay for the same days as in **a**.

Extended Data Table 1. **Real-time computation runtimes in milliseconds.** Considering the day with the highest number of trips in the year which is 43th day of the year 2011 with around 505,000 trips, we compute the breakdown of the runtimes for building a bipartite trip-vehicle graph, t_g, and finding the optimum assignment, t_{tva}, by receiving the trip requests in the next minute based on the proposed online network-based batching model. The total runtime $t_g + t_{tva}$ per batch remains under 100 milliseconds for $x = 1.2$ and 200 milliseconds for $x = 2.0$. This shows the the practicality of the proposed method from the computational point of view. We have also varied the maximum delay, Δt between 2 and 6 minutes. The average is computed for all the minutes in the day, while the max times correspond to the batch computation with the maximum runtime. The results are based on 10 separate runs for the entire day, each time reinitializing the fleet deployment warm-up phase as described in the Methods section. The experiments were performed on a Linux workstation equipped with an Intel Core i7-3930K CPU running at 3.20GHz and 32GB of RAM. For maximal fairness, the running times are based on the actual times spent running the program, not on the CPU clocks assigned to the process.

Extended Data Figure 1. **The effect of the upper bound δ on trip connection time. a**, Plot showing minimum fleet size and the vehicle void ratio as a function of the upper bound used on the trip connection time, δ. For increasing values of δ, the minimum fleet size decreases while the void ratio increases. The results are produced averaging over 14 days of data. **b**, Plot showing the yearly variation in the minimum fleet size for $\delta = 10$ and 20 minutes. This plot reports also the number of daily trips for comparison.

417

Extended Data Figure 2. | **Intra-annual patterns.** For each day in year 2011, the number of daily trips (blue line) and the corresponding minimum fleet size (green line) is shown. Both quantities obey a weekly pattern which, is, however, different: while the number of daily trips (blue line) consistently increases from Monday to Friday and drops on Sunday, the minimum number of vehicles needed to serve those trips (green line) remains substantially constant at a value around 6,200, with a considerable drop on Sundays where this value is reduced to about 4,000.

Extended Data Figure 3. | **Detailed temporal patterns of the minimum fleet.** At each time during a day, each active vehicle in the minimum fleet set operates in one of the three possible modes: 1) empty of passengers and on the way to pick up a passenger, 2) empty of passengers and waiting at a passenger's pick-up location to pick up, 3) serving with a passenger on board. The number of vehicles operating in each of these modes computed for each minute during the day follows a regular daily as well as weekly patterns as shown by the three colored curves for all months in the year. Different panels correspond to different months, and the color intensity is used to differentiate different days of the week, with the color fading towards the weekend. The total fleet-size active on the road (black-grey curves) shows robustness as most of the vehicles in the minimum fleet are active at all times during the day (see also Extended Data Fig. **4 a-b**).

Extended Data Figure 4. | **Vehicle-level performance in the minimum fleet assignment**. **a**, Stacked horizontal bar plots showing the start and end of the operation time (left and right ends of the stacked bars) for each vehicle in the minimum fleet assignment for various days in week. The day in each panel represents results for days in the second week of each month. Vehicles active times are represented by a very thin colored bar. Stacking the bars horizontally for all the vehicles in each day. The vehicles are sorted based on their start of operation time which stacking them ultimately creates each plot for each day. In all days except on an outliner day (second Saturday in March 2011), the patterns show high efficiency where majority of vehicles start early in the day and operate till the end of the day. **b**, Stacked horizontal bar plots, this time representing the total operation time of the vehicle (the length of the bar). The bars are sorted based on the vehicles total operation time, the lowest bar corresponding to the vehicle with the longest operation time. A distinct pattern emerges on most of the weekends and some days during the week. A significant percentage of vehicles on the most days of weekends operate for a short time to serve a small subset of trips which we refer to as special demand trips. We believe existence of these trips requires additional vehicles because of the way their pick-up and drop-off times and locations are distributed spatio-temporaly. **c**, The $q - q$ plot showing the percentage of trips served (vertical axis), using the vehicle-shareability minimum fleet optimization, by the percentage of vehicles represented on the horizontal axis. Vehicles are sorted based on their total operation time, i.e., the vehicles with larger operation times appear to the left of those with smaller operation times on the horizontal axis of these plots. Each panel corresponds to a day-of-week and the curves in each panel represent all such days in the entire year (e.g., all Mondays). As it can be seen on most weekends and consistent with the patterns observed in **b**, a significant percentage of vehicles (between 5-10%) serve only a very small percentage of trips ($< 1\%$). This can be observed from the appearing cusps near the top of some of the curves.

Extended Data Figure 5. | **Modeling minimum fleet size scaling with the number of trips. a**, Scatter plot showing the average operation time of vehicles in the optimal dispatching for different days versus the average number of trips per vehicle for each day. The former quantity scales linearly with the average number of trips per vehicle. This holds despite the fact that the fleet-size manifests a saturation pattern as the number of trips grow, **b**. The coefficient of proportionality between the two quantities in **a** is different and separates the weekends. The coefficient is slightly lower for Saturdays (blue) and significantly lower on Sundays (green) compared to that of weekdays. **c**, Plot showing the interplay between the minimum fleet size and the number of trips for each simulated day to manifest how the fleet size changes as the number of trips increases significantly. The supersampling is done by combining the demand in similar days in two and three successive weeks. The number of vehicles shows linear growth with a ripple-like pattern of saturation and increase as it can be seen in **c**. **d**, Plot showing the interplay between the fleet size and the number of trips as simulated using a simple bin-packing model. The oversimplified model as described in the Methods section can still capture the ripple-like saturation/increase pattern.

Extended Data Figure 6. | **Average vehicle utilization in minimum fleet assignment versus number of daily trips. a**, Scatter plot showing the average total time with a passenger on board per vehicle in a day versus the total number of daily trips for that day. Each point in the scatter plot represents a day. This quantify which is a measure of vehicle utilization in the minimum fleet assignment shows overall increasing pattern with the increase in the number of daily trips which is consistent with the fact that the minimum fleet size shows robustness. **b**, Scatter plot showing the average total wait time to pick-up passengers per vehicle versus the number of daily trips for each day. The average total wait-time decreases as the number of daily trips increases, again can be interpreted as the increase in the utilization of vehicles. The observed patterns justifies the unused capacity assumption used to develop the bin-packing model (see Methods and Extended Data Fig. **5**).

418

Extended Data Figure 7. | **Efficiency comparison between single versus multiple mobility operators**. The optimal fleet-size in the single operator versus multi-operator mobility service in each day for a sample of 100 days in year 2011. In case of multiple operators, trips are randomly assigned to one of the operators in equal proportions, and network-based optimization is performed by each operator independently. The number of vehicles needed by each operator are then summed up and the number for each operator is shown in **a**. **b**, Fleet-size percent increase plot showing how the transition from a monopolistic to a oligopolistic market incurs a drop in efficiency quantifiable in about 4 to 6% for two operators markets, and about 6 to 10% for three operators markets.The further increase in the number of operators leads to higher inefficiency in terms of fleet-size as one is moving away from the global optimum achievable in the monopolistic market to an increasingly partial one.

Extended Data Figure 8. | **Intra-annual comparison between batch and on-the-fly models**. Plots showing the percentage of trips served within the next 6 minutes from the time the trip requests are received. In the batch model the advance knowledge of the trip information is restricted to only the next minute (batch time). The trips in each batch are assigned to the available vehicles using the online version of the network approach from the minimum fleet of size 7440. This approach scores consistently higher percentage ($> 90\%$) compared to the on-the-fly model. In the on the fly model trips are assigned to the closest available vehicle. To achieve the same level of service using the on-the fly model, the fleet size must increase by more than 30% (see Fig. (**4**)**b**). The shaded region represents the 6σ variations when vehicle warm-up phase is reinitialized 50 times, where $\sigma = \max(\%) - \min(\%)$, is the difference between the percentage of served trips achieved for the runs which score maximum and minimum values for each day.

METHODS

Trip Data

The dataset used in this work consists of more than 150 million trips with passengers of all 13,586 taxicabs in New York during the calendar year of 2011. The dataset contains a number of fields from which we use the following: origin time, origin longitude, origin latitude, destination longitude, and destination latitude. The measurement precision of times is in seconds; location information has been collected by the data provider via GPS location tracking technology. Out of our control are possible biases due to urban canyons which might have slightly distorted the GPS locations during the collection process. All individual-level IDs are given in anonymized form; origin and destination values refer to the origins and destinations of trips, respectively.

Map matching

Similar to the preprocessing done in [7], we used data from www.openstreetmap.org to create the street network of Manhattan. As described in the previous work, a filtering method on the streets of Manhattan to select only the following road classes: primary, secondary, tertiary, residential, unclassified, road, and living street. We left out several other classes deliberately. These include footpaths, trunks, links, or service roads, as they are unlikely to contain delivery or pick-up locations. We extracted the street intersections to build a network in which nodes are intersections and directed links are roads connecting those intersections (we use directed links because a non-negligible fraction of streets in Manhattan are one-way). The extracted network of street intersections was then manually cleaned for obvious inconsistencies or redundancies (such as duplicate intersection points at the same geographic positions), in the end containing 4,091 nodes and 9,452 directed links. This network was used to map match the GPS locations from the trip dataset. We only matched locations for which a closest node in the street intersection network exists with a distance less than 100 m and discard the remaining trips. After the preprocessing and filtering phase more than 147 million trips remain to be used in the next phases of our analysis.

Travel times computation

Travel times information is a key part of building vehicle shareability networks. The knowledge of estimated travel times is based on a heuristic method developed and used in [7]. This method uses pick-up and drop-off times of a historical trip data set and computes the travel times between arbitrary origin/destinations on the road map.

In the following we briefly describe the core idea of this method. A detailed description can be found in the Supplementary Information presented in [7].

Each street segment belong to the set, $S = \{S_1, ..., S_h\}$, of all road segments connecting any pair of adjacent intersections in the road map. Given a set of k historical trips $\mathcal{T} = \{T_1, ..., T_k\}$, the problem of travel times computation is estimating the travel time t_i^e for each street segment $S_i \in S$ in such a way that the average relative error (computed across all trips) between the actual travel time t_i and the estimated travel time t_i^e for trip T_i computed starting from the x_i (compound with a routing algorithm) is minimized. Once error minimizing travel times for each street in S are determined, the travel time between any two intersections i and j can be computed starting from the t_is, using a routing algorithm that minimizes the travel time between any two intersections.

Following steps are involved in the process of travel times computation. First, we partition the trip set in time sliced subsets $\mathcal{T}_1, ..., \mathcal{T}_{24}$ where subset $\mathcal{T}_i$ contains all trips whose starting time is in hour i of the day. If desired, finer partitioning (e.g., per hour and weekday, per hour and weekday and month, etc.) is possible. The travel time estimation process can be performed independently on each of the time-sliced trip subsets. We define $\mathcal{T}_i^{sq}$ as the subset of trips with origin $\mathbf{x}_s$ and destination $\mathbf{x}_q$ in which $\mathbf{x}_s$ contains the (latitude, longitude) coordinates of the s-th intersection after the trips are matched. A small fraction of trips are filtered to remove "loop" trips (i.e., trips with the same origin and destination), as well as excessively "short" or "long" trips. After a step in which initial routes are computed using a pre-selected initial speed v_{int} (the same for all streets) as described in [7], a second trip filtering step is performed, in which excessively "fast" and "slow" trips are removed from the travel time estimation process. 97% of trips remain after these filtering. The travel time estimations obtained using this method are reasonable, with a relatively lower average speed of around $5.5m/sec$ estimated during rush hours (between 8am and 3pm), and peaks around $8.5m/sec$ at midnight.

Node-disjoint Path Cover

In the following we provide a set of definitions and present relevant theorems with their proofs to systematically formulate the problem of reducing the fleet-size as a path-cover problem on a vehicle-shareability network.

Given a directed network $V = (N, E)$, a path P in V is a sequence of edges $\{e_1 = (n_1^1, n_1^2), \ldots, e_k = (n_k^1, n_k^2)\} \in E$ such that $n_i^2 = n_{i+1}^1$, for each $i = 1, \ldots, k-1$. The set of nodes in path P is defined as $N(P) = \bigcup_{i=1,k}\{n_i^1\}$. The length of a path P is the number k of edges that form it.

Definition 1 (Path cover)**.** *Given a directed network $V = (N, E)$, a node-disjoint path cover of V is a collection of paths $\{P_1, \ldots, P_h\}$ such that $\bigcup_{i=1,h} N(P_i) = N$ and $N(P_i) \cap N(P_j) = \emptyset$ for any $i \neq j$. The size of the cover is the number h of paths of which it is formed.*

Note that, under the conventional assumption that zero-length paths corresponding to single nodes are allowed, a node-disjoint path cover always exists. In the following, to simplify presentation we drop the term "node-disjoint" and use "path cover" to refer to a "node disjoint path cover" as defined in Definition 1.

Theorem 1. *Let $\mathcal{C} = \{P_1, \ldots, P_h\}$ be a path cover of the vehicle shareability network $V = (N, E)$. Then, all the trips in $\mathcal{T}$ can be served by h vehicles.*

Proof. Consider a path $P = \{e_1 = (n_1^1, n_1^2), \ldots, e_k = (n_k^1, n_k^2)\}$ in the vehicle shareability network V. By definition of shareability network, the trips corresponding to n_1^1 and n_1^2 (call them T_1 and T_2) can be served by a single vehicle. Furthermore, the vehicle performing trip T_1 is guaranteed to arrive at the pick-up location of T_2 within time t_2^p; i.e., vehicle sharing does not impose any delay on the starting time of the second trip. Also, the upper bound δ on the trip connection time is not violated by definition of shareability network. Hence, the vehicle that serves T_1 and T_2 can be used to serve trip T_3 corresponding to node n_2^2 in V, since the starting time of trip T_2 is not changed due to sharing, implying that the condition ensuring shareability of T_3 and T_2 is still fulfilled. By iterating the argument across all nodes in $N(P)$, we can conclude that all trips whose corresponding nodes are in $N(P)$ can be served by a single vehicle. Thus, if a path cover of size h exists, we can conclude that all trips in $\mathcal{T}$ can be served by h vehicles. □

Corollary 1. *The minimum number of vehicles needed to serve the trips in $\mathcal{T}$ equals the size of the minimum path cover of the vehicle shareability network $V = (N, E)$.*

Finding the size of the minimum path cover of an arbitrary directed network is NP-hard [18], hence, computationally infeasible for large graphs. However, the optimal solution can be found in polynomial time if the network is *acyclic*, meaning that no there is no directed path in the network forming a closed loop.

Definition 2 (Directed Acyclic Network)**.** *A directed network $V = (N, E)$ is acyclic if it has no directed cycles, i.e., it does not contains directed paths starting at some vertex $n \in N$ and eventually returning back to n again.*

Any vehicle-shareability network as defined above is a directed acyclic network. To see how the acyclic character arises one can use proof by contradiction. Assume a cyclic path exists in V. For simplicity, assume the path has minimal length of 2. Let $P = \{(n_1, n_2), (n_2, n_1)\}$ be a cyclic path, and let T_1, T_2 be the trips corresponding to n_1 and n_2, respectively. By definition of vehicle shareability network, we have the following sequence of inequalities:

$$t_1^d \leq t_1^d + t_{12} \leq t_2^p < t_2^d \leq t_2^d + t_{21} \leq t_1^p \,,$$

which is a contradiction since $t_1^d > t_1^p$. Hence, no cyclic path of length 2 can exist in V. The proof follows by straightforwardly extending the above sequence of inequalities to cyclic paths of arbitrary length. This implies that the minimum number of vehicles needed to perform a set $\mathcal{T}$ *of trips can be computed in polynomial time*. More specifically, it is shown that for directed acyclic networks the problem of finding the path cover of minimum size is equivalent to the well-known maximum matching problem on bi-partite graphs, which can be solved in time $O(|E|\sqrt{|N|})$ using the Hopcroft-Karp algorithm.

Online model

The results shown so far compute the minimum infrastructure based on the knowledge of the entire shareability network for the considered day. This is analogous to the oracle model as defined in [7], and is consistent with a scenario in which trip requests are issued in advance (e.g., through a reservation system). To investigate to what extent the above described benefits extend to systems where trip requests are issued in real time (such as UBER, Lyft, etc.), we repeat the analysis in the so-called

online model. In the online model, we have a number of vehicles available for serving trips which is defined as $N = N_{min} \cdot x$, where N_{min} is the minimum fleet size for the day of reference as computed by the oracle model, and $x > 1$ is an inflating factor. We then start serving trip requests with the available vehicles, whose initial position is determined through a warm-up phase in which a number of trip requests from the previous day (not accounted to compute the results) are served. In order to compare online models, two possible strategies are used to dispatch vehicles and serve trip requests:

OnTheFly (OTF): trip requests are served sequentially; when a new trip request is issued, the dispatched vehicle is chosen as the first available vehicle that minimizes passenger waiting time.

Batch: trip requests are collected for time $\delta = 1$ min and processed in batches. When a batch is processed, a maximum matching is computed to maximize the number of requests that can be successfully served (i.e., served within $max(\Delta t) =$ 6mins); vehicles are then dispatched based on the result of the maximum matching algorithm as explained in the following. At each given minute the trip requests information and the location of the available vehicles are compiled to construct a weighted bipartite graph. The edge weight on a pair of vehicle-trip node represents the pick-up delay a passenger associated with the trip node would experience in case the vehicle associated with the vehicle node is chosen to serve the passenger. After constructing this weighted bipartite network, the maximum matching algorithm can be used to find a subset of edges covering maximum number of trip nodes served within the tolerable delay, $max(\Delta t)$.

Fig. **4** shows the success rate of the two dispatching algorithms for a period of 15 consecutive days, when $x = 1.2$ in serving the trips within a certain tolerable delay. As seen in this figure, the batch method (blue lines) provides a success rate which is consistently above 92 percent, and much higher than what achieved by the sequential OTF method for $max(\Delta t) = 6$mins. As reported in Extended Data Table **1**, the running times of the online version of the method are below $200\ msec$ in the worse case scenarios on a standard Linux machine, indicating feasibility of the proposed approach for real-time optimization.

The warm-up phase used in the above mentioned online optimizations consists of first deploying each vehicle at a random intersection, then running the batch optimization algorithm as described above on the 2 hours of historical trip requests that precede the period of interest. The shaded regions in Fig. **4a,c** and Extended Data Fig. **8** represent the variation in the percentage of the trip served as obtained by running the real-time optimization for each day multiple times, each time reinitializing the warm-up phase with a distinct random initial deployment of the fleet. The variations are quite small showing that within two hours the system's spatio-temporal distribution does not depend significantly on the initial deployment after a two-hour warm-up.

Limiting the degrees of node connectivity in vehicle-shareability network via trip connection time constraint

We defined the vehicle shareability network in such a way that nodes which represent individual trips are connected if only it is feasible for a vehicle to serve those trips one after the other without introducing any delay in their pick-up and drop-off times. Checking whether two trips satisfy such criteria requires the knowledge of travel times in the city which is estimated using the method described previously. Since this network definition puts no constraints on the connectivity apart from the consecutively serving feasibility, the number of network links grows fast with the increase in the number of trips. This is due to the fact that trips separated by a large enough time gap between their drop-off and pick-up times can always be served with the same vehicle even though they may be spatially far from each other. This leads to a very high connectivity in the vehicle shareability network as most pairs of trips separated enough temporally can satisfy such connectivity condition. To limit the number of edges in the network, and to make sure that the vehicles do not operate without any passenger onboard for too long leading to underutilization and increase in the void ratio as the fraction of time vehicles operate without a passenger, we introduce an upper bound on the connection time between the trips. The connection time is defined as the time a vehicle operates without a passenger between the consecutive trips.

The first issue to address is how to set the bound δ on the trip connection time, which is a parameter that can be used to tradeoff fleet size with vehicle and traffic efficiency. On one hand, when δ is decreased to 0 we approach a situation in which each trip is served by a dedicated vehicle: a solution with maximal vehicle utilization which is optimal also for traffic (if we assume that vehicles somehow appear at the origin and disappear at the destination of the trip they serve), but incurring prohibitive costs for the mobility operator. On the other hand, when δ grows excessively the fleet size is reduced, however at the expense of decrease in the operational and traffic efficiency since some vehicles may be on the road for long times without any passenger on board between serving the trips. Thus, how to set δ is an important design choice that shall be left in the hands of mobility operators, traffic authorities and policy makers.

Extended Data Fig. **1** shows how we come up with a reasonable setting for δ. The plot reports both the minimum fleet size as well as the average fraction of time a vehicle spends for connecting consecutive trips (void ratio) in seconds, for increasing values of δ. As expected, the former quantity decreases with δ, while the latter increases. For values of δ larger than 15 min, however, the vehicle fleet-size decreases only marginally, while the void ratio still increases. For this reason, for the results reported in the main text we have set $\delta = 15$ min. For reference, the right panel of Extended Data Fig. **1** reports the yearly analysis of minimum fleet size – similar to what reported in Extended Data Fig. **2** – for $\delta = 10$ and 20 min.

Vehicle utilization

Better understanding of the efficiency of the network-based vehicle-trip assignment requires having a closer look into the patterns of the utilization of the individual vehicles in the minimum fleet. The overall time each vehicle spends during its operation in a day consists of traveling with a passenger on board, without any passenger and on the way to pick up the next one, or waiting at the pick-up location of a new passenger. Ultimately, the goal in an efficient vehicle-trip assignment is to maximize overall utilization while minimizing the operation costs. This can be achieved for each vehicle when the fraction of time a vehicle operates without a passenger on board is minimized.

Extended Data Fig. **2** reports the yearly analysis of minimum fleet requirements, along with corresponding daily number of trips. While the number of daily trips clearly displays an increasing weekly pattern, the number of required vehicles remains fairly constant, with a dip on Sundays. The robustness of the fleet-size despite the significant variation in the robustness shows that minimum fleet can tolerate handling extra trips which their addition does not require new vehicles. Addition of such trips certainly leads to higher vehicle utilization as we show here.

Extended Data Fig. **3** reports a breakdown of the deployed vehicles into the different phases of deployment – passenger onboard, en-route to next passenger, waiting for next passenger – for a better understanding of the utilization patterns.

Extended Data Fig. **4** reports the vehicle-level performance using various temporal metrics. The vehicle start and end of operation time during the day in Fig. **4a** shows that for the majority of the days, minimum fleet assignment leads to high operation times for the majority of the vehicles.

The reported plots in Extended Data Fig. **4 b-c** on some days clearly show existence of a small fraction of under-utilized vehicles operating on average for less than two hours, serving what we called "special-purpose" trips. These trips mostly occur on the weekend and are spatio-temporaly isolated, meaning that their existence requires new vehicles as the existing vehicle-trip assignment cannot be rearranged to successfully accommodate these trips.

A bin-packing model to describe the fleet-size scaling

As shown in the Extended Data Fig. **5**, for a large number of days with daily trips ranging from 350,000 to 550,000, there is only a small variation in the minimum fleet size. This pattern seems a bit counter-intuitive at first glance as basic logic implies that an increase in the number of trips should somehow lead to increase in fleet-size. Outside this range this expected increasing pattern holds and for smaller number of trips we have more or less a linear scaling (see the result of supersampling in Fig. **5c**).

To explain the saturation pattern observed in Fig. **2**, we come up with a simple bin-packing model to show that the reason for fleet-size robustness in a certain range is related to an existing spatiotemporal capacity/tolerance to accommodate more trips in the minimum fleet. Consider a set of N vehicles with a fixed spatio-temporal capacity to accommodate k trips during a day. The exact value of k depends on the average duration of a trip during a given time of the day, and it is limited by an strict upper bound equal to 24 hours on the maximum vehicle operation time. We start with a configuration where we have a certain number of trips $N \cdot x$ ($x << k$) randomly distributed in the bins with a poisson distribution. We start to add one trip at a time and randomly sample a small subset of n vehicles as candidate set. (n is a hyper parameter of the model that we assume to be either 1 or 2). Two scenarios can happen: 1) a subset of the selected vehicles still have capacity to accommodate more trips: In that case we randomly select one of them and assign the trip to that vehicle. 2) none of the vehicles have spatio-temporal capacity to accommodate the new trip. In that case we add a new vehicle to the system to accommodate the new trip. We repeat this process and model the relationship between the number of vehicles and number of trips in this manner.

The interesting plateau-increase pattern emerges from this model which implies that for some intermediate ranges, the fleet-size first increases and then shows some robustness with respect to further increase in the number of trips consistent with the observed pattern as observed based on our minimum fleet optimization approach in Fig. **2**. This simple model suggests that the reason for the minimum fleet-size robustness is that the probability of finding a vehicle which can successfully accommodate that new trip is still significant as many cars operate with significant unused spatiotemporal capacity when the number of trips is relatively low. The range of minimum fleet-size tolerance is determined by the maximum number of trips that a certain number of vehicles can serve in theory. This maximum number depends on the spatio-temporal distribution of trips, especially the distribution of the trip durations. For instance if the average trip duration in a day is 10-15 minutes, a vehicle can serve up to around 3-4 trips per hour assuming a 5 minute connection time between the trips on average. This way the upper bound would be around 100 trips for vehicles that are active for most of 24 hours. With this assumption the maximum number of trips a minimum fleet of around 6,000 vehicles can tolerate is around 600,000 trips. Fig. **2** and the results of the model in Fig. **5d** support this argument.

Although the model in this section is an oversimplification and does not consider the complex spatiotemporal constraints in whether a vehicle can serve a trip, it, however, does capture the saturation pattern represented in Fig. **2**. Extended Data Fig. **6** supports the idea that the robustness of the fleet-size is due to the existing capacity in vehicles by showing how the metrics

associated with the vehicle utilization show consistent increase in the vehicle utilization for days with higher number of trips. Days with higher number of trips score higher average utilization per vehicle as it can be seen in both, the increase in the average time a vehicle spends on the road with a passenger on board for each day (see Extended Data Fig. **6a**) and also the increase in the average time vehicles spend waiting to pick up a passenger at the pick-up point (see Extended Data Fig. **6b**).

Multi-operator model

As briefly discussed in the main text, consider a situation in which there are more than one mobility operators, each having access only to a subset of trip demand data. Assuming that the operators assign the vehicles in their fleet to trip demand they have access to without sharing information with the other mobility operators. The question is to what extent the fleet-size is affected by the lack of information sharing between a certain number of mobility operators which is equivalent of going from a global optimum towards a local one where each vehicle receives a limited information about adjacent trips and tries to maximize its utilization independent of other vehicles. Of course the latter is the extreme limit where the number of operators is very large and only a local optimum can be achieved. In the following using a simplified model we try to address the case for two and three mobility operators equality sharing the mobility market.

For this purpose we randomly sample from the trip demand data at each given point in time and divide the trip set into multiple subsets. For each trip subset we can build a vehicle shareability network and do the minimum fleet optimization as described in this Letter. Each optimization leads to a minimum fleet size for each mobility operator. By comparing the sum of the fleet sizes for the multiple mobility operator with the global minimum fleet-size as described in the Letter we can find out how much away we are from the global optimum.

Extended Data Fig. **7** shows the temporal pattern of the sum of fleet-sizes for a sample of 100 days from NYC taxi trip data. To obtain the right estimation for the sum of fleet-sizes, we have divided trip set in each day into two and three equally sized subsets by random subsampling. We have repeated the random subsampling several times and each time we perform the vehicle-shareability network optimization to find the fleet-size for each subset. The average of the fleet-sizes obtained from several random subsampling for each day is then presented in Extended Data Fig. **7a-b**. As shown in Extended Data Fig. **7b**, the transition from a monopolistic to a oligopolistic market incurs a small drop in efficiency quantifiable in about 4 to 6 percent for two operators markets, and about 6 to 10 percent for three operators markets. The further increase in the number of operators leads to higher inefficiency in terms of fleet-size as one is moving away from the global optimum achievable in the monopolistic market to an increasingly partial one. One can see that if the number of disjoint operators increases further the total size of the fleet would keep increasing due to the lack of communication between mobility operators even in the case when each of them try to optimize their fleet-size based on the information about the trip demand they receive. The fact that considering two or three operators sharing equal shares of the mobility market results only in a small drop in efficiency in terms of the fleet-size, shows that the minimum fleet size optimization using the network based approach for two or three independent operators is not far from the global optimum.