LETTER

https://doi.org/10.1038/s41586-018-0095-1

Addressing the minimum fleet problem in on-demand urban mobility

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

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 best to size and operate a fleet of vehicles, given a certain demand for personal mobility. Previous studies[1–5] either do not 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), of how to determine 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 dataset 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 per cent reduction in fleet size compared to current taxi operation. Although constraints on driver availability and the existence of abnormal trip demands may lead to a relatively larger optimal value for the fleet size than that predicted here, the fleet size remains robust for a wide range of variations in historical trip demand. These predicted reductions in fleet size follow directly from a reorganization 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 towards 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 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. 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 the 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[17] are obtained using a model of transportation based on aggregated mobility data and Euclidean spatial assumptions, and hence lack the resolution necessary to estimate the urban-scale benefits of vehicle sharing accurately.

We start from the notion of the shareability network introduced in ref. [7], which did not focus on the dispatching of vehicles. The type of shareability network introduced here is profoundly different from the type studied previously: it models the sharing of vehicles, whereas previous networks[7–9] modelled the sharing of rides. The main methodological contribution of this Letter is to show 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. Besides revealing a structural property of vehicle-sharing networks, this connection allows the derivation of computationally efficient algorithms for optimal vehicle deployment and dispatching. Although 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 $(t_i^{\mathrm{p}},\ t_i^{\mathrm{d}},\ l_i^{\mathrm{p}},\ l_i^{\mathrm{d}})$ 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. 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, calculated using a travel-time estimation model and assuming the passenger leaves the pick-up location at time t_i^{p}. In contrast to ref. [17], travel times here are computed using the actual road network, and using global positioning system (GPS)-based estimations derived from the taxi trip dataset that account for hourly variations in traffic, as in ref. [7]. If the set $\mathcal{T}$ is extracted from a real-world dataset (for example, taxi trips), the times t_i^{p} and t_i^{d} represent the actual times 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

[1]Senseable City Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Istituto di Informatica e Telematica del CNR, Pisa, Italy. [3]Department of Mathematics, Cornell University, Ithaca, NY, USA. [4]Cornell Tech, New York City, NY, USA. [5]These authors contributed equally: M. M. Vazifeh, P. Santi. *e-mail: mmvazifeh@gmail.com

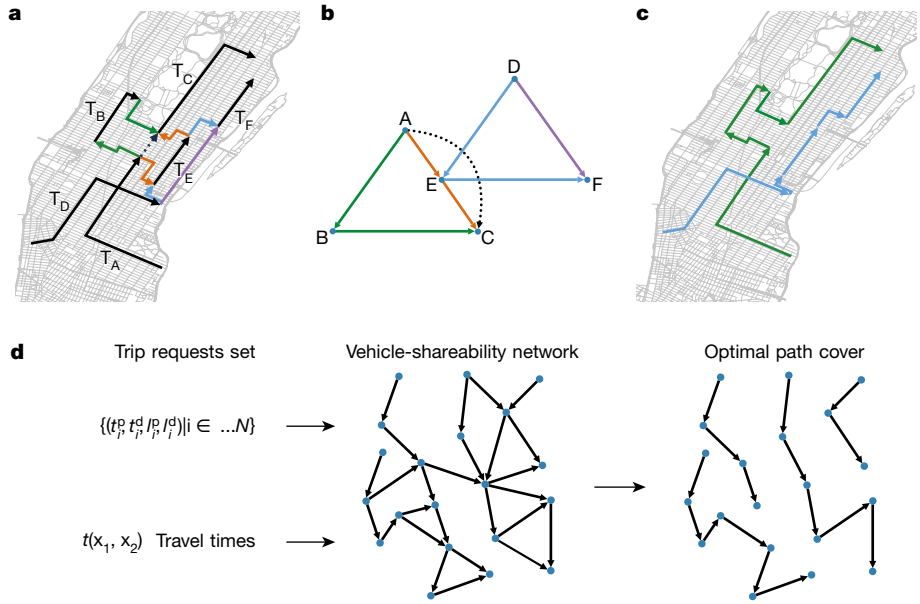


Fig. 1 | Constructing the vehicle-shareability network. a, Several trip requests are depicted on the map index as $T_A,\ldots, T_F$. The coloured path on the map represents different possibilities for vehicle dispatching. **b**, Colours are used in the vehicle-shareability network to specify how various dispatches can be represented using paths on this network. One of the two dispatching scenarios required only two vehicles whereas 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 there is a large enough gap in time between the drop-off of the first and the pick-up point of the next trip 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.

vehicle is guaranteed to be available at the desired location and time. On the other hand, the above problem formulation might entail substantial inefficiencies for the operator and the environment. Consider two consecutive trips T_A and 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 travelled 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 (1) a vehicle is available at each l_i^p on or before t_i^p and (2) the connection time is at most δ minutes', where the upper bound δ on the connection time is a problem parameter.

Figure 1 illustrates the construction of the vehicle-shareability network that enables the minimum fleet problem to be optimally solved 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 (1) above) and $t_j^p - t_i^d \leq \delta$ (which accounts for assumption (2) above). 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—that is, a dispatch. 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, that is, 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 bipartite 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 for formal proofs.

We have tested our methodology on a dataset of over 150 million taxi trips performed in the city of New York in the year 2011. This dataset has been selected from a number of available datasets[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 compare our methodology directly with current taxi operation. The data have been sliced into daily datasets $\mathcal{T}_i$, 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 that is also optimal 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, the 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 is left to mobility operators, traffic authorities and policy makers. In this study, we set $\delta = 15$ min, as explained in Methods. The results of our method with different values of δ are reported in Methods (see Extended Data Fig. 1).

Figure 2 shows the daily number of vehicles needed to address the entire taxi demand in New York City 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 (inset to Fig. 2a) 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 spatiotemporal patterns are likely to play a large part as well. To investigate this issue further, we have analysed daily vehicle usage in the optimal solution.

The vehicle usage analysis reported in Methods 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 probably related to the extra nightly demand. The analysis also highlighted clear weekly patterns in

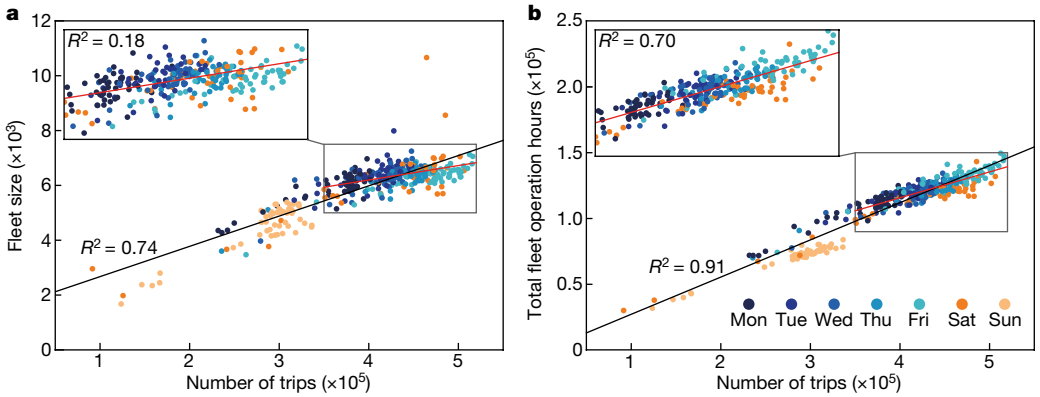


Fig. 2 | Minimum fleet-size analysis. a, The interplay between the number of daily trips and the minimum number of vehicles required to serve them. The colours of the dots correspond to different weekdays. Clustering of points with the same colour suggests weekly patterns, which are confirmed by the yearly analysis reported in Methods (see Extended Data Fig. 2). The plot shows a moderate correlation between the two quantities. However, when focusing on days with a 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 of 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, estimated by summing over the operation times for each vehicle in the minimum fleet obtained over 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. The total fleet operation hours have a much stronger correlation with the number of trips than with the number of vehicles.

vehicle use, 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 at weekends hints also at an opportunity to perform routine vehicle maintenance on a weekly basis.

A better scaling law relating vehicle fleet size to the daily number of trips can be obtained by defining a metric for fleet sizing that incorporates how long a vehicle is used during a day. We define a 'full-time equivalent' vehicle as a vehicle continuously operating 24 h a day. (In the case of human-driven vehicles, we can think of having the vehicle operated in three 8-h shifts, for instance.) Figure 2b shows that the scaling law relating the number of daily trips with full-time equivalent vehicles is more accurate than the previous one, with the coefficient of determination R^2 value increased from 0.74 to 0.91, and from 0.18 to 0.70 for the trip-intense days reported in the inset.

Figure 3 shows the efficiency breakthrough provided by network-based optimization: when compared to current taxi operation in New York City, 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 refs. [7,9]. That fleet size can be reduced by as much as 40% without the use of ride sharing and with no delay for passengers has, to the best of our knowledge, not 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

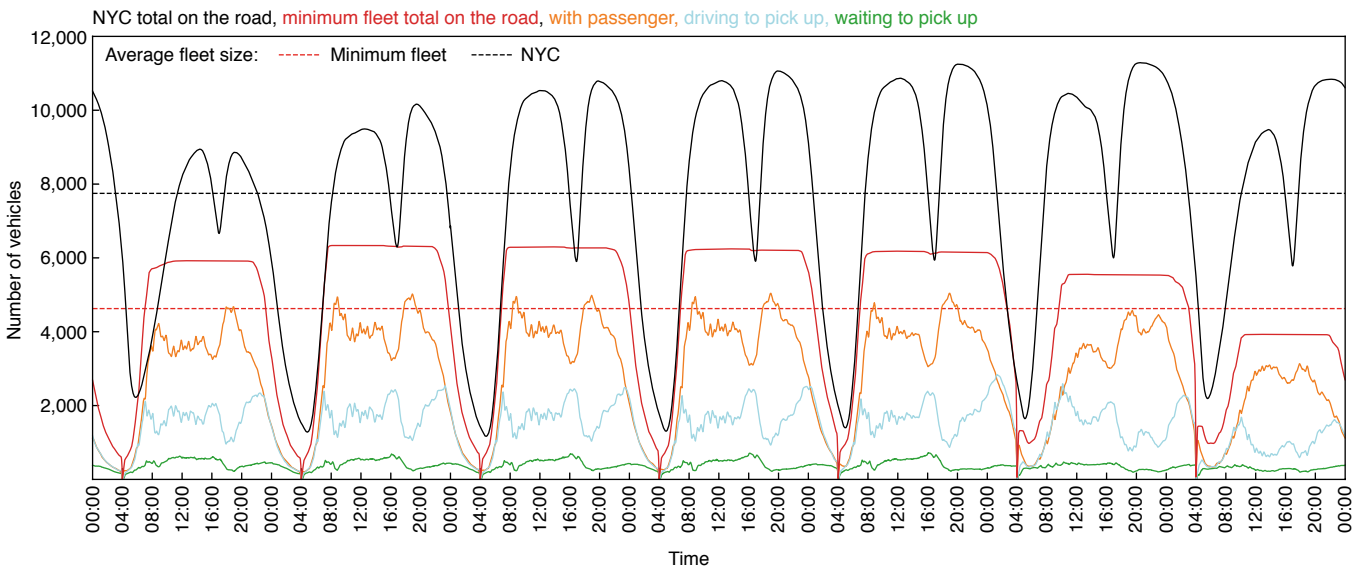


Fig. 3 | Fleet efficiency comparison. Comparison between actual number of New York City (NYC) taxis on the road[6] (black curve), versus the minimum number of vehicles as computed by our optimal approach (red curve). For the optimized approach, a breakdown of vehicles into 'with passenger', 'waiting to pick up' and 'driving to pick up' is also reported. The curves are computed by 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 to implement using a smartphone app, our method provides a near-optimal 30% reduction in the number of circulating taxis (see Fig. 4).

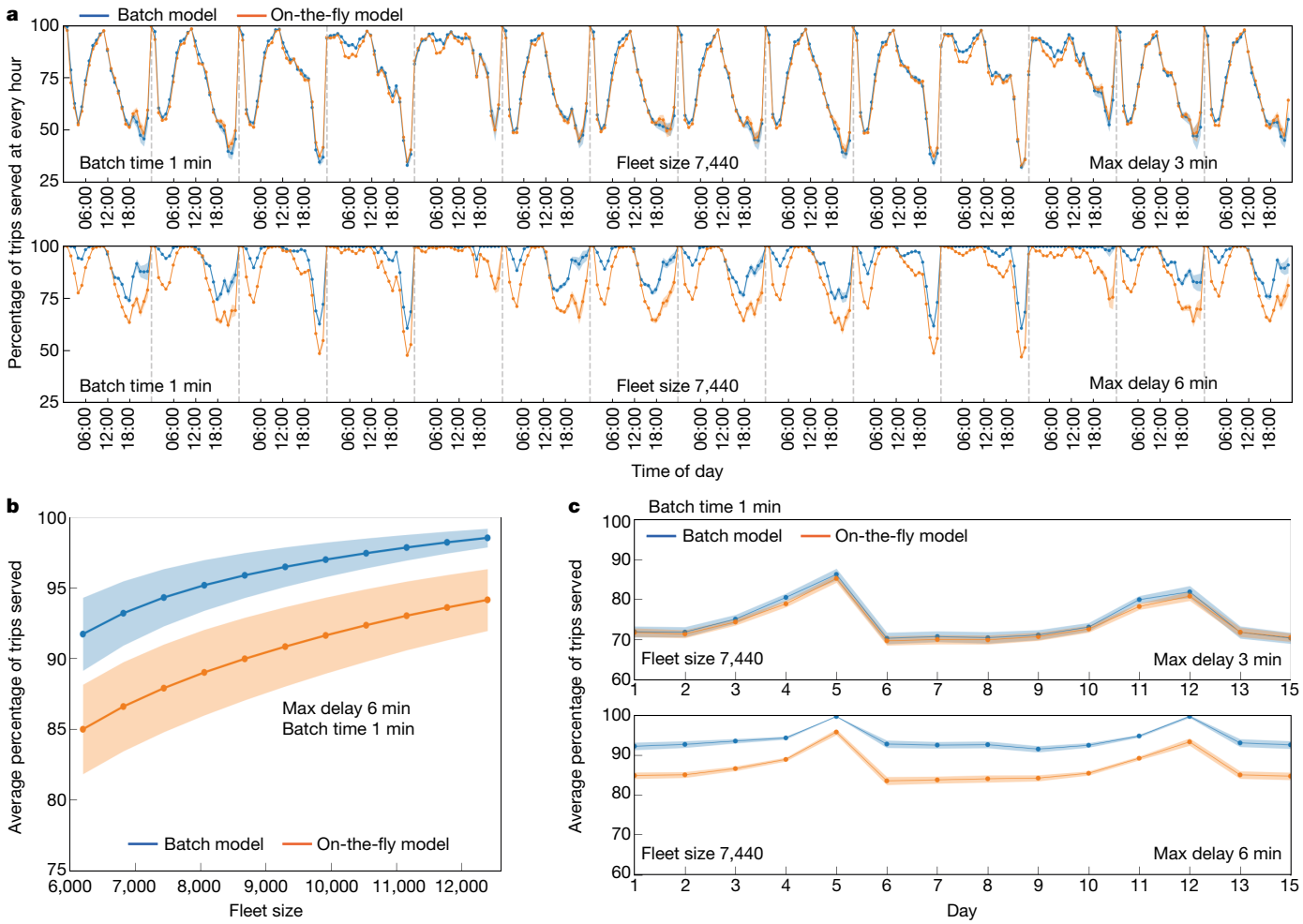


Fig. 4 | Performance of the network-based online vehicle-dispatching model. a, Two-panel plot comparing the percentage of trips served within a certain maximum delay $\max(\Delta t)$ based on the batch-based optimized dispatching (blue line) versus sequential dispatching, as described in the Methods. The panel at the top represents the results for $\max(\Delta t) = 3$ min and the one at the bottom is for $\max(\Delta t) = 6$ min. The use of network-based dispatching (blue 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 running the algorithm with daily trip knowledge 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 $\max(\Delta t) = 3$ min or 6 min, and the batching time in batch-based dispatching is set to 1 min. **b**, The average percentage of trips successfully served within a 6-min delay is above 90% for a fleet of 6,200 vehicles. Similar performance can only be achieved with 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**.

with the online version of the algorithm reported in the Methods. This version collects trip requests for a short time, for example, one minute, and locally optimizes vehicle dispatching based on this limited knowledge. Figure 4 shows that, with a 30% fleet reduction and using the online version of the algorithm, 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 trip requests and vehicle-dispatching decisions are centralized, a model that 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-theory notion of the 'price of anarchy'[20] 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. Although optimal from the vehicle operation and environmental viewpoint, a monopolistic market is however highly undesirable for many other reasons, most importantly, lack of competition with consequent higher prices for customers. An additional analysis reported in Methods shows that most of the efficiency benefits of centralized vehicle operation are still possible in an oligopolistic market.

Although the characterization of minimum fleet size reported here 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 produce relatively larger values of the minimum fleet requirement than those predicted here. Extending the concept of the vehicle-sharing network to incorporate driver constraints is possible and is left for further analysis.

Broader effects on traffic are foreseen if our methodology is to be used for optimizing urban 'on-demand' mobility services more in general, especially in a future of autonomous vehicles. However, it is well known that an improvement in mobility efficiency is 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 we leave to future work.

Finally, we observe that, while applied here to taxi trips as a case study, the proposed methodology for optimal vehicle fleet sizing and dispatching is general and can be applied to model any type of point-to-point mobility. However, the approach presented here focuses on optimizing and dispatching 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 here for optimal solution of the 'minimum fleet' problem.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0095-1.

Received: 19 July 2017; Accepted: 21 February 2018;
Published online 23 May 2018.

1. Berbeglia, G., Cordeau, J. F. & Laporte, G. Dynamic pick-up and delivery problems. *Eur. J. Oper. Res.* **202**, 8–15 (2010).
2. Laporte, G. The vehicle routing problem: an overview of exact and approximate algorithms. *Eur. J. Oper. Res.* **59**, 345–358 (1992).
3. Baker, B. M. & Ayechew, M. A. A genetic algorithm for the vehicle routing problem. *Comput. Oper. Res.* **30**, 787–800 (2003).
4. Yang, J., Jaillet, P. & Mahmassani, H. Real-time multivehicle truckload pick-up and delivery problems. *Transport. Sci.* **38**, 135–148 (2004).
5. Clare, G. L. & Richards, A. G. Optimization of taxiway routing and runway scheduling. *Proc. IEEE Intell. Transport. Syst.* **12**(4), 1000–1013 (2011).
6. Bloomberg, M. R. & Yassky, D. *New York City Taxicab Factbook* http://www.nyc.gov/html/tlc/downloads/pdf/2014_tlc_factbook.pdf (New York City Taxi and Limousine Commission, 2014).
7. Santi, P. et al. Quantifying the benefits of vehicle pooling with shareability networks. *Proc. Natl Acad. Sci. USA* **111**, 13290–13294 (2014).
8. Tachet, R. et al. Scaling law of urban ride sharing. *Sci. Rep.* **7**, 42868 (2017).
9. Alonso-Mora, J., Samaranayake, S., Wallar, A., Frazzoli, E. & Rus, D. On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment. *Proc. Natl Acad. Sci. USA* **114**, 462–467 (2017).
10. Rosenberg, T. It's not just nice to share, it's the future. *NY Times Opinionator* https://perma.cc/89YT-VHVF (2013).
11. Sundararajan, A. From Zipcar to the sharing economy. *Harvard Business Review* https://hbr.org/2013/01/from-zipcar-to-the-sharing-eco (2013).
12. Mitchell, W. J., Borroni-Bird, C. E. & Burns, L. D. *Reinventing the Automobile: Personal Urban Mobility for the 21st Century* (MIT Press, Cambridge, 2010).
13. Botsman, R. & Rogers, R. *What's Mine is Yours: The Rise of Collaborative Consumption* (HarperCollins, New York, 2010).
14. Handke, V. & Jonuschat, H. *Flexible Ridesharing* (Springer, Berlin, 2013).
15. United Nations Environment Programme *UNEP 2010 Annual Report*; https://www.unenvironment.org/resources/annual-report/unep-2010-annual-report (UNEP, Nairobi, 2011).
16. Barth, M. & Boriboonsomsin, K. Real-world carbon dioxide impacts of traffic congestion. *Transp. Res. Rec.* **2058**, 163–171 (2011).
17. Spieser, K. et al. in *Road Vehicle Automation* (eds Meyer, G. & Sven Beiker, S.) 229–245 (Springer, 2014).
18. Boesch, F. T. & Gimpel, J. F. Covering the points of a digraph with point-disjoint paths and its application to code optimization. *J. Assoc. Comput. Mach.* **24**, 192–198 (1977).
19. Hopcroft, J. & Karp, R. An $n^{(5)/(2)}$ algorithm for maximum matching in bipartite graphs. *SIAM J. Comput.* **2**, 225–231 (1973).
20. Roughgarden, T. *Selfish Routing and the Price of Anarchy* (MIT Press, Cambridge, 2005).

Acknowledgements P.S., M.M.V. and C.R. thank all sponsors and partners of the MIT Senseable City Laboratory including Allianz, the Amsterdam Institute for Advanced Metropolitan Solutions, the Fraunhofer Institute, Kuwait-MIT Center for Natural Resources and the Environment, Singapore-MIT Alliance for Research and Technology (SMART) and all the members of the Consortium. The research of S.H.S. was supported by USA NSF grant numbers DMS-1513179 and CCF-1522054.

Reviewer information *Nature* thanks G. Casey, V. Knoop and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions P.S. defined the problem, designed the solution and algorithms, and contributed to the analysis and paper writing. M.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 helped with the analysis. S.H.S. contributed to writing. C.R. supervised the research and contributed to writing.

Competing interests The authors declare no competing interests.

Additional information
Extended data. is available for this paper at https://doi.org/10.1038/s41586-018-0095-1.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0095-1.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.M.V.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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 City 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 (that is, streets with high-rise buildings on both sides), which might have slightly distorted the GPS locations during the collection process. All individual-level identifications 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 ref. [7], we used data from www.openstreetmap.org to create the street network of Manhattan. As described in previous work[7], we used 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 matched only GPS locations (both trip origins and destinations) that are within 100 m of at least one node in the street intersection network, which is the case for the majority of trips, and discarded the remaining ones. After the preprocessing and filtering phase, more than 147 million trips remain to be used in the next phases of our analysis.

Travel-time computation. Travel-time 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 ref. [7]. This method uses pick-up and drop-off times of a historical trip dataset and computes the travel times between arbitrary origins and 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 of ref. [7].

Each street segment belongs to the set, $S=\{S_1,\ldots,S_h\}$, of all road segments connecting any pair of adjacent intersections in the road map. Given a set of k historical trips $\mathcal{T}=\{T_1,\ldots,T_k\}$, the problem of travel-time 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 routeing 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_i values, using a routeing algorithm that minimizes the travel time between any two intersections.

The following steps are involved in the process of travel-time computation. First, we partition the trip set in time-sliced subsets $\mathcal{T}_1,\ldots,\mathcal{T}_{24}$ where subset $\mathcal{T}_i$ contains all trips whose starting time is in hour i of the day. If desired, finer partitioning (for example, per hour and weekday, per hour and weekday and month, and so on) 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 x_s and destination x_q in which x_s contains the (latitude, longitude) coordinates of the sth intersection after the trips are matched. A small fraction of trips are filtered to remove 'loop' trips (that is, 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 ref. [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 this filtering. The travel-time estimations obtained using this method are reasonable, with a relatively lower average speed of around 5.5 m s^{-1} estimated during rush hours (between 8 am and 3 pm), and peaks at around 8.5 m s^{-1} 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-covering 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 of edges k 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 $\cup_{i=1,h}N(P_i)=N$ and $N(P_i)\cap N(P_j)=\varnothing$ for any $i\neq j$. The size of the cover is the number of paths h of which it is formed.

We 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 (Path cover)' above.

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; that is, 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 the 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 as a result of 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] and hence computationally infeasible for large graphs. However, the optimal solution can be found in polynomial time if the network is acyclic, meaning that 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, that is, it does not contains directed paths starting at some vertex $n\in N$ and eventually returning 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 and T_2 be the trips corresponding to n_1 and n_2, respectively. By the definition of vehicle-shareability network, we have the following sequence of inequalities:

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

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 bipartite graphs, which can be solved in time $O(|E|(|N|)^{1/2})$ using the Hopcroft–Karp algorithm.

Online model. The results shown so far compute the minimum infrastructure on the basis of the knowledge of the entire shareability network for the day considered. This is analogous to the Oracle model as defined in ref. [7], and is consistent with a scenario in which trip requests are issued in advance (for example, 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 and Lyft), 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 $N=N_{\min}x$, 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. To compare online models, two possible strategies are used to dispatch vehicles and serve trip requests, as follows.

On-the-fly. 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 (that is, served within max-$(\Delta t)=6$ min); vehicles are then dispatched on the basis of the result of the maximum matching algorithm, as explained in the following. At each given minute the trip requests information and the locations 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 the maximum number of trip nodes served within the tolerable delay, $\max(\Delta t)$.

Figure 4 shows the success rate of the two dispatching algorithms for a period of 15 consecutive days, for $x = 1.2$, 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%, and is much higher than what is achieved by the sequential on-the-fly method for $\max(\Delta t) = 6$ min. As reported in Extended Data Table 1, the running times of the online version of the method are below 200 ms in the worst-case scenarios on a standard Linux machine, indicating the 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 h of historical trip requests that precede the period of interest. The shaded regions in Fig. 4a and c and in Extended Data Fig. 3 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 2 h the system's spatiotemporal distribution does not depend noticeably on the initial deployment.

Limiting node connectivity via trip connection time. We defined the vehicle-shareability network in such a way that nodes that represent individual trips are connected only if 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 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 feasibility of consecutively serving trips, the number of network links grows quickly with the increase in the number of trips. This is because trips separated by a large enough time gap between their drop-off and pick-up times can always be served by the same vehicle although they may be spatially far from each other. This leads to a very high connectivity in the vehicle-shareability network because most pairs of trips separated by enough time can satisfy this 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 an increase in the void ratio (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 trade off fleet size against vehicle and traffic efficiency. On the 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 maximum vehicle utilization that is also optimal 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, but this is at the expense of a decrease in the operational and traffic efficiency because 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, which should 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 connecting consecutive trips (the 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, whereas 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 is reported in Extended Data Fig. 2—for $\delta = 10$ min and 20 min.

Vehicle utilization. A better understanding of the efficiency of the network-based vehicle-trip assignment requires 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 travelling 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 is 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 the corresponding daily number of trips. Whereas 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 large variation in the number of daily trips shows that the minimum fleet size can tolerate handling extra trips without needing extra vehicles. The addition of such trips certainly leads to higher vehicle utilization, as we show here. Extended Data Fig. 4 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. 5 reports vehicle-level performance using various temporal metrics. The vehicle start and end of operation time during the day in Extended Data Fig. 5a shows that on most days, minimum fleet assignment leads to high operation times for the majority of the vehicles. The reported plots in Extended Data Fig. 5b and c on some days clearly show the existence of a small fraction of under-used vehicles operating on average for less than two hours, serving what we call 'special-purpose' trips. These trips occur mostly on the weekend and are spatiotemporally isolated, meaning that their existence requires new vehicles because the existing vehicle–trip assignment cannot be rearranged to accommodate these trips successfully.

A bin-packing model to describe fleet-size scaling. As shown in Extended Data Fig. 6, 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, because 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 a smaller number of trips we have a more-or-less linear scaling (see the result of supersampling in Extended Data Fig. 6c).

To explain the saturation pattern observed in Fig. 2, we use a simple bin-packing model to show that the reason for fleet-size robustness within a certain range is related to an existing spatiotemporal capacity to accommodate more trips in the minimum fleet. Consider a set of N vehicles with a fixed spatiotemporal 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 h on the maximum vehicle operation time. We start with a configuration where we have a certain number of trips Nx (where $x \ll k$) 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 hyperparameter of the model that we assume to be either 1 or 2). Two scenarios are possible: (1) a subset of the selected vehicles still have the capacity to accommodate more trips, in which case we randomly select one of them and assign the trip to that vehicle; (2) none of the vehicles have spatiotemporal capacity to accommodate the new trip, in which 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.

An interesting plateau-then-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 a further increase in the number of trips, consistent with the observed pattern as observed from 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 relatively high as many cars operate with a large 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 spatiotemporal distribution of trips, especially the distribution of the trip durations. For instance, if the average trip duration in a day is 10–15 min, a vehicle can serve up to around 3–4 trips per hour assuming a 5-min connection time between the trips on average. In this way the upper bound would be around 100 trips for vehicles that are active for most of the day. With this assumption the maximum number of trips a minimum fleet of around 6,000 vehicles can tolerate is around 600,000 trips. Figure 2 and the results of the model in Extended Data Fig. 6d support this argument.

Although the model in this section is an oversimplification and does not consider the complex spatiotemporal constraints that determine whether a vehicle can serve a trip, it does, however, capture the saturation pattern represented in Fig. 2. Extended Data Fig. 7 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 vehicle utilization show a consistent increase in vehicle utilization for days with higher numbers of trips. Days with higher numbers of trips score higher average utilization per vehicle as 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. 7a) and also in the increase in the average time vehicles spend waiting to pick up a passenger at the pick-up point (see Extended Data Fig. 7b).

Multi-operator model. As briefly discussed in the main text, consider a situation in which there is more than one mobility operator, each having access only to a subset of trip demand data and assuming that the operators assign the vehicles in their fleet to the trip demands they have access to without sharing information

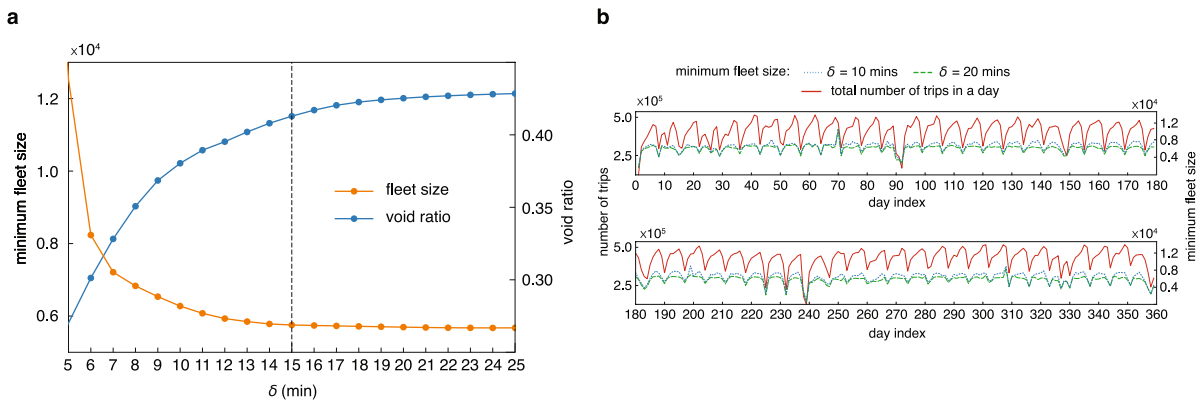
with the other mobility operators. The question is to what extent the fleet size is affected by the lack of information sharing between a certain number of mobility operators. This is equivalent to going from a global optimum to a local optimum, in which each vehicle receives limited information about adjacent trips and tries to maximize its utilization independent of other vehicles. This latter situation is the extreme limit at which 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 cases for two and three mobility operators equally sharing the mobility market.

For this purpose we randomly sample 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 case with the global minimum fleet size we can find out how far away we are from the global optimum.

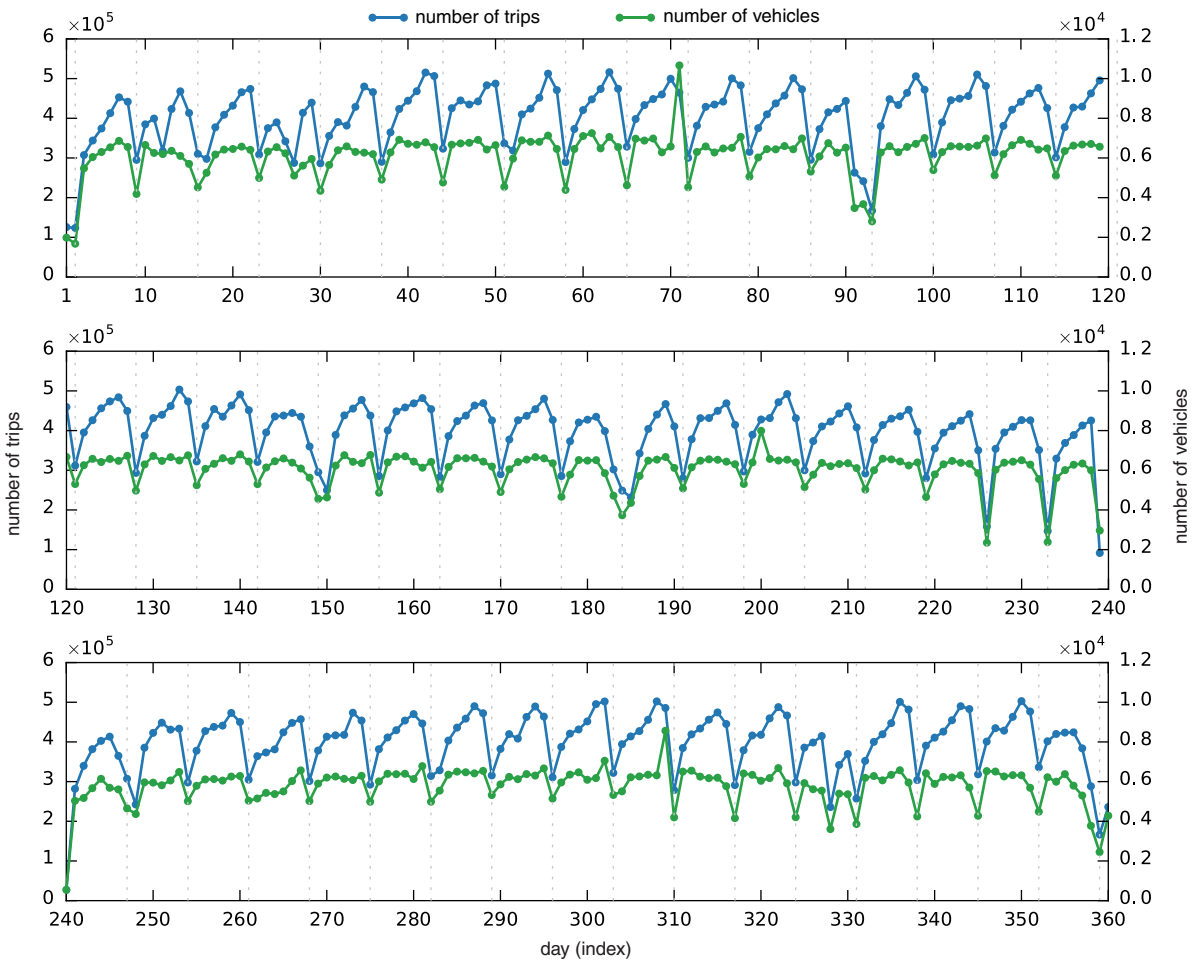
Extended Data Fig. 8 shows the temporal pattern of the sum of fleet sizes for a sample of 100 days from New York City taxi trip data. To obtain a good estimate for the sum of fleet sizes, we have divided the trip set in each day into two and three equally sized subsets by random subsampling. We repeated 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 fleet size obtained from several random subsamplings each day is then presented in Extended Data Fig. 8a and b. As shown in Extended Data Fig. 8b, the transition from a monopolistic to a oligopolistic market incurs a small drop in efficiency quantifiable at about 4%–6% for two-operator markets, and about 6%–10% for three-operator markets. A 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. If the number of disjoint operators increases further the total size of the fleet would keep increasing owing 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 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.

Fleet-size inflation due to rare events. The analysis of historical data has shown that our model provides a robust improvement (the reduction in fleet size) on previous models. However, an inflation of the optimal fleet size could occur as a result of rare, unusual demand patterns. For instance, if a sudden burst in the number of trips occurs around a given location but with diverging destinations, these trips will not be connected to each other in the vehicle-shareability network. Thus, for such cases to be served, the trips require separate vehicles from the existing pool on the road or even extra vehicles. These special events inflate the number of vehicles required because the nodes added to the vehicle-shareability network can have sparse or no connectivity to other nodes in the network. Although rare, such cases of trip-demand bursts can occur after events such as sports matches or concerts. However, based on our historical analysis it is evident that these outlying patterns only rarely lead to any inflation in the number of vehicles required to serve all the demand.

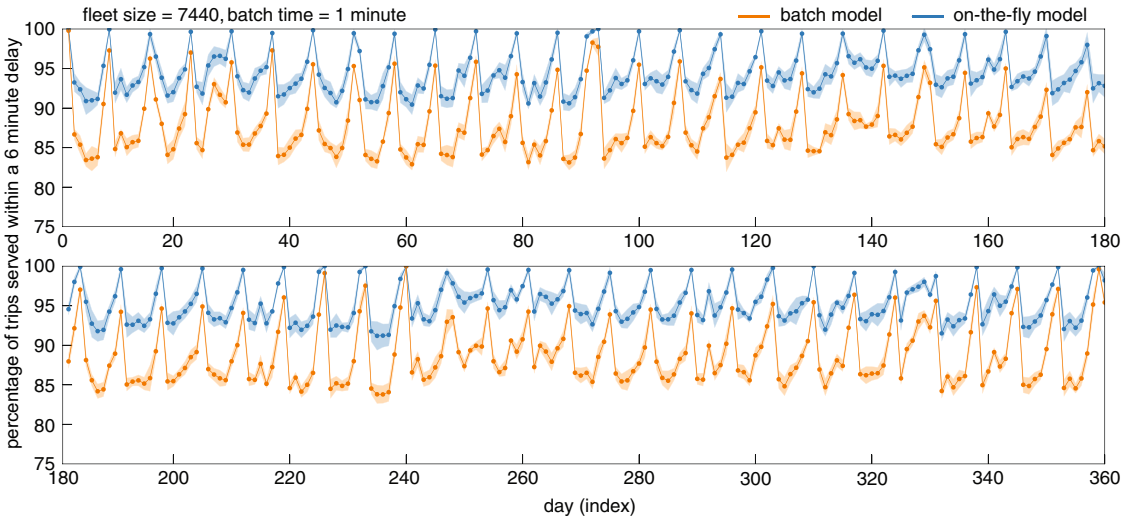
Data and code availability. All data processed during the course of this study are included in this Letter and its Supplementary Information. The code for generating the shareability networks and optimal dispatching is subject to licensing and could be made available upon request to the authors. New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



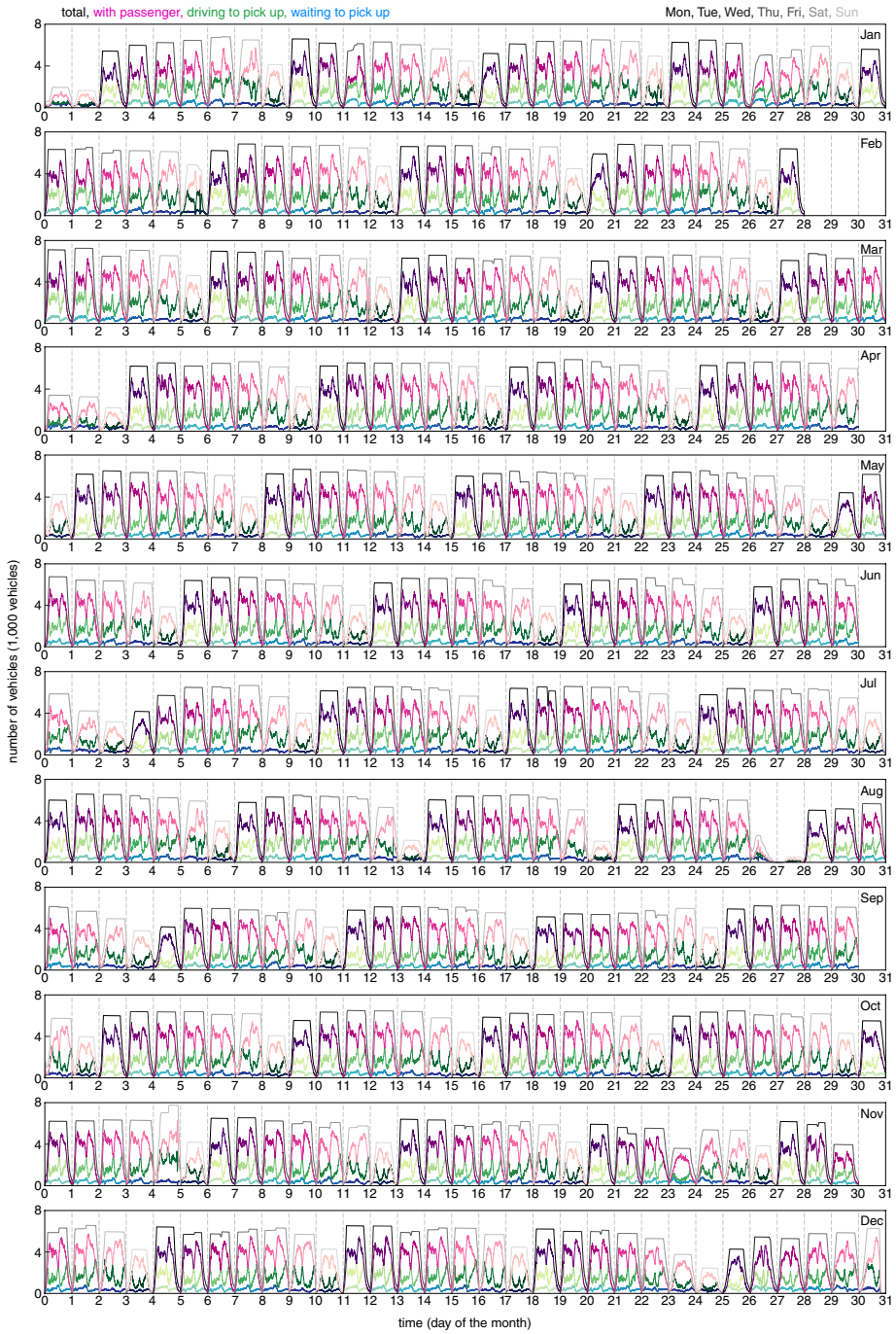
Extended Data Fig. 1 | The effect of the upper bound δ on trip connection time. a, Plot showing the minimum fleet size and the vehicle void ratio as a function of the upper bound used on the trip connection time, δ. For increasing values of δ, the minimum fleet size decreases while the void ratio increases.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



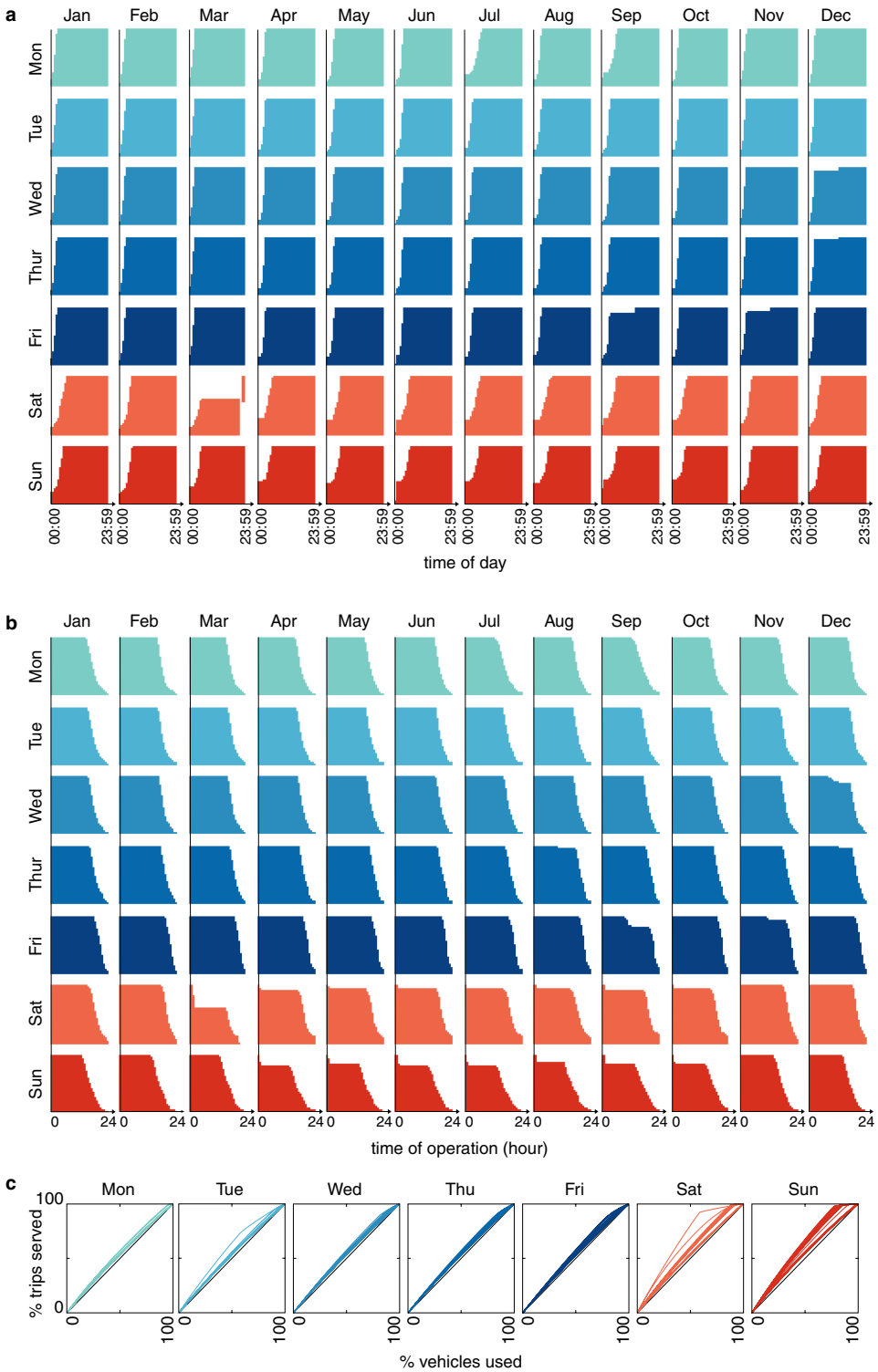
Extended Data Fig. 2 | Intra-annual patterns. For each day in year 2011, the number of daily trips (blue line) and the corresponding minimum fleet size (green line) are shown. Both quantities follow different weekly patterns: whereas the number of daily trips (blue line) consistently increases from Monday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



Extended Data Fig. 3 | Intra-annual comparison between batch and on-the-fly models. Plots showing the percentage of trips served within the next 6 min from the time the trip requests are received. In the batch model, 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 7,440. This approach scores a consistently higher percentage (90%) than does 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. 4b). The shaded region represents the 6σ variations when the 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 that score maximum and minimum values for each day.

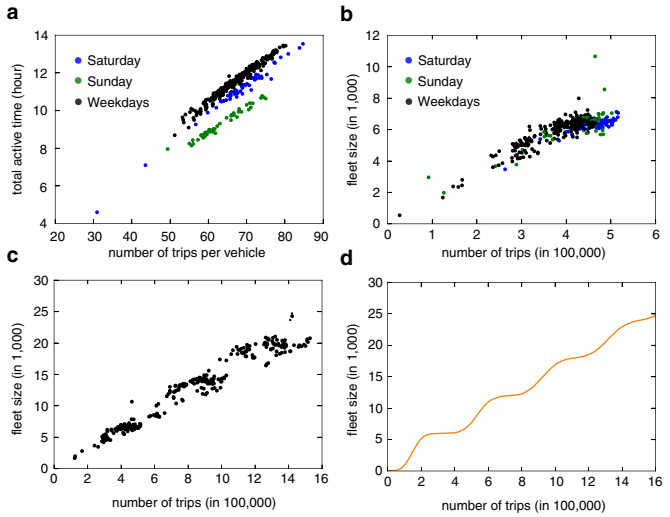


Extended Data Fig. 4 | Detailed temporal patterns of the minimum fleet. At each time during a day, each active vehicle in the minimum fleet set operates in one of three possible modes: (1) empty of passengers and on the way to pick up a passenger, (2) empty of passengers and waiting at a passenger's 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 regular daily and weekly patterns, as shown by the three coloured curves for all months in the year. The total fleet size active on the road (black-to-grey curves) demonstrates robustness, because most of the vehicles in the minimum fleet are active at all times during the day (see also Extended Data Fig. 5a and b). Different panels correspond to different months, and the colour intensity is used to differentiate different days of the week.

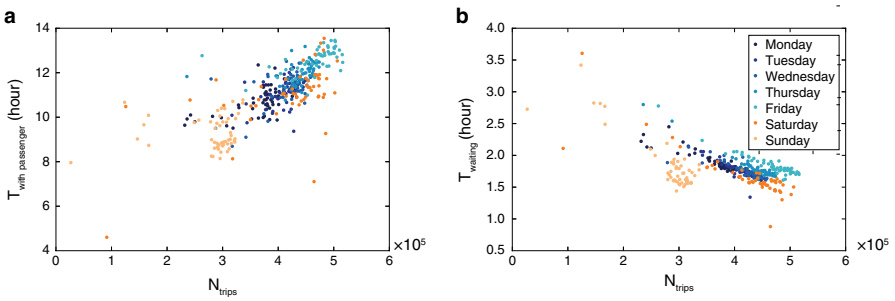


Extended Data Fig. 5 | See next page for caption.

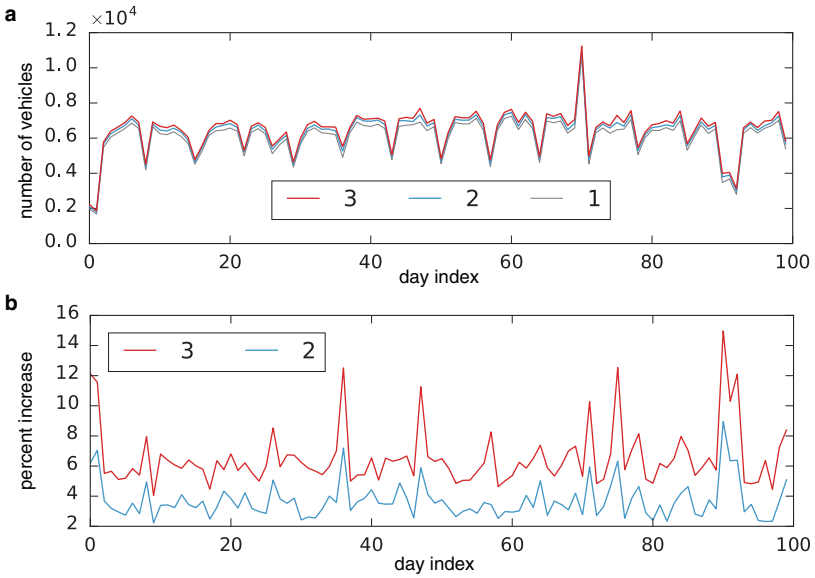
Extended Data Fig. 5 | Vehicle-level performance in the minimum fleet assignment. a, Stacked horizontal bar plots showing the start and end of the operation time (left and right ends of the stacked bars) for each vehicle in the minimum fleet assignment for various days of the week. The day in each panel represents results for days in the second week of each month. Vehicle active times are represented by a very thin coloured bar. The vehicles are sorted by the start of operation time, and stacking them horizontally creates each plot for each day. In all days (except for the outlier day of the second Saturday in March 2011), the patterns show high efficiency, with the majority of vehicles starting early in the day and operating until 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 vehicle's 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 on some days during the week. A substantial percentage of vehicles on most weekends operate for a short time to serve a small subset of trips, which we refer to as special-demand trips. We believe that the existence of these trips requires additional vehicles because of the way their pick-up and drop-off times and locations are distributed spatiotemporally. **c**, The $q-q$ plot (where the quantile is the fraction of points below the given value) showing the percentage of trips served (vertical axis), using the vehicle-shareability minimum fleet optimization, with the percentage of vehicles represented on the horizontal axis. Vehicles are sorted on the basis of their total operation time, that is, the vehicles with longer operation times appear to the left of those with shorter operation times on the horizontal axis of these plots. Each panel corresponds to a day of the week and the curves in each panel represent all such days in the entire year (for example, all Mondays). On most weekends and consistent with the patterns observed in **b**, a large percentage of vehicles (between 5% and 10%) serve only a very small percentage of trips ($<1\%$). This can be observed from the cusps appearing near the top of some of the curves.



Extended Data Fig. 6 | Modelling minimum fleet size scaling with the number of trips. a, Scatter plot showing the average operation time of vehicles with 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 out the weekends. The coefficient is slightly lower for Saturdays (blue) and much 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 greatly increases. The supersampling is done by combining the demand for similar days in two and three successive weeks. The number of vehicles shows linear growth with a ripple-like pattern of saturation and increase. **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 described in Methods can still capture the ripple-like saturation/increase pattern.



Extended Data Fig. 7 | Average vehicle utilization in the minimum fleet assignment versus the 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. The average total time with a passenger on board, which is a measure of vehicle utilization in the minimum fleet assignment, shows an overall increasing pattern with the increase in the number of daily trips that is consistent with the fact that the minimum fleet size shows robustness. **b**, Scatter plot showing the average total waiting time to pick up passengers per vehicle versus the number of daily trips for each day. The average total waiting time decreases as the number of daily trips increases, which again can be interpreted as an increase in the utilization of vehicles. The observed patterns justify the unused capacity assumption used to develop the bin-packing model (see Methods and Extended Data Fig. 6).



Extended Data Fig. 8 | Efficiency comparison between single and multiple mobility operators. The optimal fleet size in the single-operator and the multi-operator mobility service in each day for the first 100 days (1 January corresponds to day index 1) in the year 2011. In the 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 and the number for each operator is shown in **a**. **b**, Fleet-size percentage increase plot showing how the transition from a monopolistic to a oligopolistic market incurs a drop in efficiency of 4%–6% for a two-operator market, and of about 6%–10% for a three-operator market. The further increase in the number of operators leads to higher inefficiency in terms of fleet size as it moves away from the global optimum achievable in the monopolistic market to an increasingly fragmented market.

Extended Data Table 1 | Real-time computation run times in milliseconds

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

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