scientific reports

Check for updates

OPEN

Collective dynamics of capacity-constrained ride-pooling fleets

Robin M. Zech[1], Nora Molkenthin[2], Marc Timme[1] & Malte Schröder[1]✉

Ride-pooling (or ride-sharing) services combine trips of multiple customers along similar routes into a single vehicle. The collective dynamics of the fleet of ride-pooling vehicles fundamentally underlies the efficiency of these services. In simplified models, the common features of these dynamics give rise to scaling laws of the efficiency that are valid across a wide range of street networks and demand settings. However, it is unclear how constraints of the vehicle fleet impact such scaling laws. Here, we map the collective dynamics of capacity-constrained ride-pooling fleets to services with unlimited passenger capacity and identify an effective fleet size of available vehicles as the relevant scaling parameter characterizing the dynamics. Exploiting this mapping, we generalize the scaling laws of ride-pooling efficiency to capacity-constrained fleets. We approximate the scaling function with a queueing theoretical analysis of the dynamics in a minimal model system, thereby enabling mean-field predictions of required fleet sizes in more complex settings. These results may help to transfer insights from existing ride-pooling services to new settings or service locations.

Human mobility is a quintessential example of a complex system[1,2]. Interactions of individual travelers with each other, with their environment or with transportation services give rise to complex emergent mobility patterns and collective dynamics[2–6]. Statistical physics approaches have helped to reveal universal patterns in the scaling of human mobility[2,7], characterize recurring aspects of the structure of mobility and transportation networks[8–12], and explain fundamental properties of congestion and its persistence across a variety of systems[3,13–17]. Currently, human mobility is transforming towards new modes of transport that are increasingly self-organized and networked[2,4–6,18]. In particular, app-based on-demand ride-pooling services promise to reduce the economic and ecological impact of congestion and emissions in urban mobility, especially in light of the current trend of ongoing urbanization[19–22].

By combining trips of passengers along the same direction, ride-pooling reduces the required number of vehicles and the total distance driven. Similar to standard ride-hailing, on-demand ride-pooling services typically act as door-to-door transport for passengers, matching similar passenger requests to each other or to vehicles already on route, ideally without any detour for the passengers (Fig. 1a,b). In contrast to ride-hailing services, however, the assignment of passenger requests to ride-pooling vehicles is much more complex[23,24] due to the restrictions of the routes of the vehicles by already assigned passengers. The resulting complex collective dynamics of the ride-pooling fleet[25,26] and the intricate dependence of the service efficiency on the system parameters[23,27,28] are far from fully understood. Previous studies have analyzed the potential to pair passenger requests as a graph covering problem[23] and demonstrated a universal scaling of the theoretical potential to combine rides with similar origin and destination across empirical demand patterns from different cities[27]. Recently, similar scaling laws have been demonstrated also in a simplified dynamical model of ride-pooling in the special case of unlimited passenger capacity[25]. However, similar to restrictions from already accepted requests, capacity limits of ride-pooling vehicles constrain the assignment of new requests to vehicles. A request that cannot be served by a vehicle due to capacity constraints must be picked up and delivered by another vehicle, potentially causing route changes and additional delays (see Fig. 1c). Thus, even this simple constraint on individual vehicles may strongly affect the collective dynamics of the ride-pooling fleet as a whole and thereby also change the dynamic scaling laws.

Here, we analyze the collective dynamics of ride-pooling fleets under capacity constraints and identify the effective number of vehicles available to serve a request as the relevant scaling parameter to characterize their efficiency. With this effective available fleet size, we map the dynamics of capacity-constrained ride-pooling fleets to an unconstrained system, generalizing the scaling laws of ride-pooling efficiency. Moreover, we develop

[1]Center for Advancing Electronics Dresden (cfaed) and Institute for Theoretical Physics, Technische Universität Dresden, 01062 Dresden, Germany. [2]Complexity Science, Potsdam Institute for Climate Impact Research, 14473 Potsdam, Germany. ✉email: malte.schroeder@tu-dresden.de

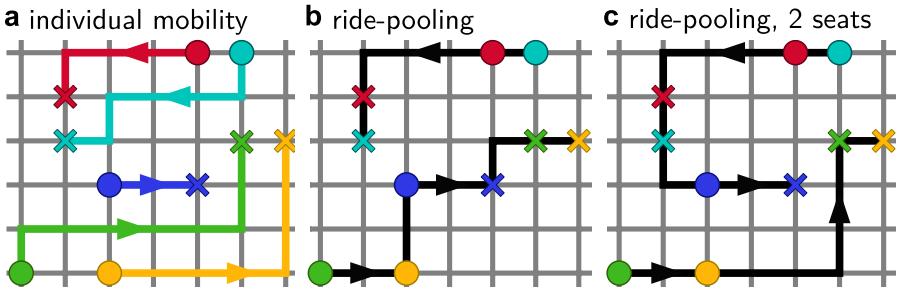

Figure 1. Constraints shape the dynamics of ride-pooling. (**a**) With individual mobility, each person travels from their origin (circle) to their destination (cross) using their own car (colored lines). (**b**) Ride-pooling combines trips along similar routes into the same vehicle. Two vehicles (black lines) starting at the green and cyan origin, respectively, serve all requests. (**c**) Constraints modify the dynamics of the ride-pooling service. If only two customers can be transported by each vehicle at a time, the dark blue trip cannot be served as in panel (**b**). Instead, the routes of the vehicles are modified and the customer is delayed.

a queueing theory description of the ride-pooling dynamics in a minimal model system that enables an approximate analytical calculation of the efficiency and the relevant scaling parameters. Together with a self-consistent mean-field approximation in more complex settings, we demonstrate the possibility of using the scaling law to estimate required fleet sizes. Overall, our results suggest that universal scaling laws of ride-pooling efficiency may hold across a much broader range of settings and constraints and may thus enable the a-priori optimization of ride-pooling fleet size, capacity, and other system parameters in previously unserviced areas.

Results

Collective dynamics of ride-pooling. The dynamics of the ride-pooling fleet depend on a large number of system parameters. The topology of the underlying street network $\mathcal{G}$ and the demand distribution ρ in space determine the average trip distance $\langle l \rangle$ across all requests. The demand distribution in time, characterized by the average request rate λ, determines the number of requests. The number of vehicles B and their properties, such as the typical velocity v or passenger capacity θ, as well as the dispatcher algorithm $\mathcal{A}$, assigning requests to vehicles, critically determine the resulting routes of the vehicles and thereby the service quality.

We simulate the dynamics of the ride-pooling service in a simplified model. Customers request transport from one node of the underlying street network $\mathcal{G}$ to another node uniformly randomly following a Poisson process with rate λ. Each request is immediately assigned to a vehicle, adjusting its planned route, such that the request is delivered as fast as possible without delaying previous requests or exceeding the capacity constraints of the vehicles. Over time, vehicles drive along their planned routes, picking up and dropping off passengers, and the system settles into a steady operating state such that the average number $\langle C \rangle$ of scheduled requests per vehicle (on board or scheduled to be picked up in the future) becomes constant if the system does not overload (Fig. 2a). We simulate these dynamics on various different network topologies, including simple network structures such as a minimal two-node graph or a complete graph, effectively one-dimensional topolgies in cycle graphs, as well as two-dimensional square lattices and geometric random networks. A more detailed description of the ride-pooling model and simulation parameters is provided in the Methods.

To compare the dynamics across different settings, we define the normalized load[25]

$$x = \frac{\lambda \langle l \rangle}{vB}, \tag{1}$$

describing the total average requested trip distance $\lambda \langle l \rangle$ per time relative to the maximal distance vB that all vehicles can drive. The load x is a lower bound for the average occupancy of the ride-pooling vehicles. When $x > 1$, more distance is requested from the system than the vehicles can drive and ride-pooling is necessary to serve all requests. Stable operation of a ride-pooling service with maximum passenger capacity θ per vehicle is, in principle, possible for loads $x < \theta$. The service necessarily overloads for $x > \theta$ since each vehicle would need to transport more than θ customers on average to serve all requests.

Capacity-unconstrained ride-pooling efficiency. The efficiency of a ride-pooling service can be consistently quantified across different settings based on the collective dynamics of the ride-pooling fleet[25]. If the capacity constraints of the system are sufficient to serve all requests, the system settles into a steady operating state with a constant number $\langle C \rangle$ of scheduled requests per vehicle (Fig. 2a). The exact value of $\langle C \rangle$ depends on the underlying network topology and system parameters (Fig. 2b). Under ideal conditions, requests are picked up immediately and delivered on the direct route to their destination. In this optimal service limit, each vehicle transports exactly x passengers on average. The average number of scheduled requests per vehicle is equal to the average occupancy and equal to the normalized load, $\langle C \rangle_{\text{opt}} = \langle O \rangle_{\text{opt}} = x$. The actual number of scheduled requests $\langle C \rangle$ in a given system is typically larger since customers may have to wait for pickup or may be subject to

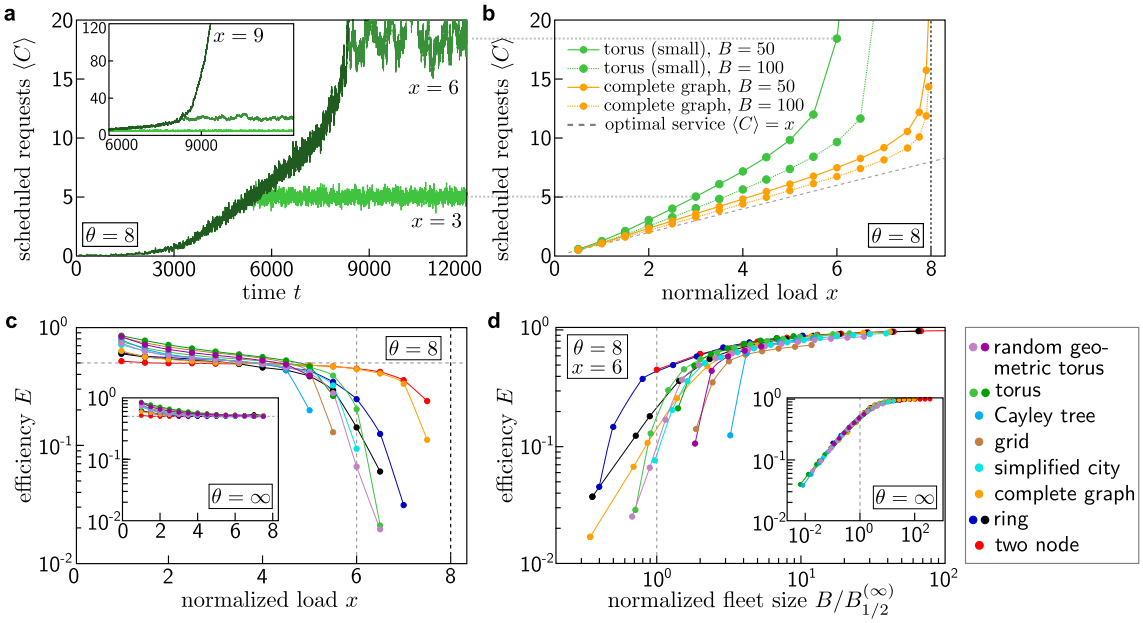

Figure 2. Capacity constraints break the topological universality of ride-pooling efficiency. (**a**) The average number $\langle C\rangle$ of scheduled customers per vehicle settles into a steady state for $x < \theta$ when the normalized load x [Eq. (1)] is slowly increased. If the normalized load is larger than the capacity, $x > \theta$, the system overloads and the number of scheduled customers increases indefinitely (inset). (**b**) For small loads, the average number of scheduled customers per vehicle increases approximately linearly with the normalized load x. The difference to the best possible scaling $\langle C\rangle = x$ (dashed line) quantifies the efficiency of the service (see panels (**c**) and (**d**)). When the load x approaches the capacity limit θ, the number of scheduled customers diverges as the system overloads. (**c**) Capacity-constrained systems behave qualitatively differently across network topologies when the load x approaches the capacity limit $\theta = 8$ of the system. (inset) Systems with unlimited vehicle capacity converge to the same efficiency E for large loads x. Fleet sizes in both simulations are identical and chosen such that the efficiency E of the capacity unconstrained systems (inset) converges to $E = 1/2$. (**d**) The efficiency curves $E_{\mathscr{A}}(\mathscr{T}, B, \theta, x)$ of the capacity-constrained systems reveal strong differences between the various network topologies (colors), especially in settings with small fleet sizes. Neither the normalized topological factor $B_{1/2}^{(\infty)}(\mathscr{T})$ nor a load-dependent scaling factor $B_{1/2}(\mathscr{T}, x)$ is sufficient to recover the topological universality observed for capacity-unconstrained systems [inset, Eq. (5)]. Colors represent different underlying networks, see "Methods" for details on the settings and simulations.

detours in the pooled rides. The difference of the number of scheduled requests $\langle C\rangle$ with respect to the optimal service limit thus quantifies the efficiency (Fig. 2c,d) of the ride-pooling system as[25]

$$E = \frac{x}{\langle C\rangle} \in [0, 1]\,. \tag{2}$$

In general, fewer vehicles or a higher request rate, i.e. an increasing normalized load x, reduce the efficiency of a ride-pooling system as more requests have to be served with fewer vehicles in the same amount of time, resulting in longer waiting times and potential detours. However, a system with higher request rate λ and more vehicles B (keeping the normalized load x constant) operates closer to the perfect service limit. More vehicles increase the options for assigning requests while the increased request rate results in more similar requests that can be easily pooled, thus adding fewer constraints per request to the routing problem (Fig. 2b,[23,25,27]). Importantly, the system efficiency E as defined above is directly related to the average service time $\langle \Delta t_s\rangle$ from the perspective of customers. During the average service time $\langle \Delta t_s\rangle$ of a single customer, a vehicle cycles on average exactly once through all its scheduled customers, i.e. dropping off all $\langle C\rangle$ customers that were scheduled earlier. During this time, a total of $\lambda\ \langle \Delta t_s\rangle$ requests are made to the system on average, of which a fraction $1/B$ is assigned to a specific vehicle. In the steady operating state, the average number of scheduled customers is thus given by

$$\langle C\rangle = \frac{\lambda\ \langle \Delta t_s\rangle}{B}\,. \tag{3}$$

Using Eqs. (1) and (2), the efficiency

$$E = \frac{x}{\langle C\rangle} = \frac{xB}{\lambda\ \langle \Delta t_s\rangle} = \frac{\langle l\rangle}{v}\,\frac{1}{\langle \Delta t_s\rangle} \tag{4}$$

thus also quantifies the service efficiency from the customer perspective[25].

The resulting efficiency $E_{\mathscr{A}}(\mathscr{T}, B, x, \theta)$ of a ride-pooling system with dispatcher $\mathscr{A}$ is a function of an effective topology $\mathscr{T} = (\mathscr{G}, \rho)$ that combines the street network topology with the spatial demand distribution, the fleet size B, the normalized load x, and the capacity θ of the vehicles. For ride-pooling systems with unlimited capacity $\theta = \infty$, this efficiency follows a universal scaling function $f_{\mathscr{A}}$,

$$E_{\mathscr{A}}(\mathscr{T}, B, x, \infty) = f_{\mathscr{A}}\left(\frac{B}{B_{1/2}(\mathscr{T}, x)}\right), \tag{5}$$

with a single scaling parameter $B_{1/2}(\mathscr{T}, x)$ summarizing the effect of the topology and the demand distribution[25]. For sufficiently large loads $x > 1$ in the ride-pooling regime, the scaling parameter $B_{1/2}(\mathscr{T}, x)$ becomes approximately constant and we replace it with a single value $B_{1/2}^{(\infty)}(\mathscr{T})$ for each effective topology $\mathscr{T}$ (Fig. 2d inset).

However, systems that behave similarly without a capacity limit, exhibit stark differences in their efficiencies after introducing capacity constraints (Fig. 2c,d). The capacity constraints seem to break the universality, especially as the system load approaches the capacity limit, $x \to \theta$ (Fig. 2c). In contrast to the capacity unconstrained systems (Fig. 2d inset, Eq. (5)[25]), the resulting efficiency curves for the capacity-constrained systems do not collapse (Fig. 2d). For fixed values of x and θ we find that the scaling is qualitatively different across topologies.

Capacity-constrained ride-pooling efficiency. Can we recover the topological universality under capacity constraints and, if so, which are the relevant scaling parameters?

To understand the effect of the capacity constraints on the ride-pooling efficiency we examine their impact on the vehicle dynamics. The pick up and delivery dynamics along a planned route of a vehicle remain unchanged for capacity-constrained systems as the route of a vehicle is planned with respect to its capacity (i.e. all planned pick-ups are always possible). The capacity constraints thus only affect the routes and the fleet dynamics by modifying the assignment of requests.

Consider a system with a large fleet size and high efficiency. When a new request arrives, only vehicles that could serve the request with almost no delay are relevant options for the assignment (Fig. 3a). In both the capacity-constrained and unconstrained system, this excludes vehicles far away from the origin of the request. Similarly, vehicles close to the origin whose currently planned route is incompatible with the request are excluded since assigning the request to them would result in unfeasibly long waiting times or detours. Compared to the unconstrained system, capacity constraints further limit the pool of feasible options by excluding vehicles that would exceed their capacity constraints during the trip, thus resulting in longer delays. The dynamic routing decision effectively becomes identical to that of an unconstrained system without those unfeasible vehicles.

Assuming a homogeneous distribution of the unavailable, fully occupied vehicles among the pool of vehicles offering the most efficient trips, this argument suggests that the capacity-constrained system behaves similarly to a capacity unconstrained system with a reduced effective fleet size

$$B_{\text{eff}}(\mathscr{T}, B, x, \theta) = \left[1 - p_{\text{delay}}(\mathscr{T}, B, x, \theta)\right] B\,. \tag{6}$$

This effective available fleet size characterizes the change in collective dynamics of the ride-pooling service due to capacity constraints. Consequently, the efficiency $E_\theta(B)$ of the capacity-constrained system is similar to the efficiency $E_\infty(B_{\text{eff}})$ of an unconstrained system with the reduced fleet size B_{eff} (Fig. 3b). To quantify the fraction p_{delay} of unavailable vehicles, we measure the probability that the optimal assignment for a request is not possible due to the capacity constraints, i.e. the request is delayed compared to the capacity unconstrained system.

This relation between capacity-constrained and -unconstrained ride-pooling dynamics suggests that the topological universality observed in unconstrained systems extends to capacity-constrained systems with the same scaling parameter $B_{1/2}$ and the effective fleet size B_{eff} (or equivalently the average fraction p_{delay} of unavailable vehicles) as a second scaling parameter. Figure 3c illustrates the collapse of the efficiency curves to a generalized universal scaling function

$$E_{\mathscr{A}}(\mathscr{T}, B, x, \theta) = f_{\mathscr{A}}\left(\frac{B_{\text{eff}}(\mathscr{T}, B, x, \theta)}{B_{1/2}(\mathscr{T}, x)}\right) \tag{7}$$

of a single parameter with $B_{\text{eff}} = (1 - p_{\text{delay}})\, B$, recovering the scaling of the unlimited capacity system with $p_{\text{delay}} = 0$ (Fig. 3c). In contrast to the scaling parameter $B_{1/2}$ describing the topological universality, the effective fleet size B_{eff} depends on all system parameters, $(\mathscr{T}, B, x, \theta)$.

This scaling relation holds even for systems operating under high loads up to large values of $p_{\text{delay}} \lesssim 0.8$. In systems operating very close to the capacity limit with $p_{\text{delay}} \to 1$ and possibly $B_{\text{eff}} < 1$, this mapping to a capacity unconstrained system begins to break down as also vehicles far away from the origin or with large detours become relevant for the assignment. These deviations are more likely for systems with strongly limited vehicle capacity or with very few vehicles.

Mean-field queueing theory predictions. Analytical calculations in a minimal two-node model confirm our results. With two nodes at a distance $\langle l \rangle$, vehicles travel back and forth between the nodes without detours for customers. A vehicle arrives at a single node every $2\langle l \rangle / (vB)$ time units on average. From the point of view of the node, all vehicles are identical since they always drop off all current customers when arriving and transport up to θ customers requesting a trip from that node. If vehicles are distributed equidistantly and never idle, the queueing dynamics at each node effectively follows a queue with Poisson distributed requests, a deterministic service interval $2\langle l \rangle / (vB)$ with batch service for at most θ customers at the same time, and a

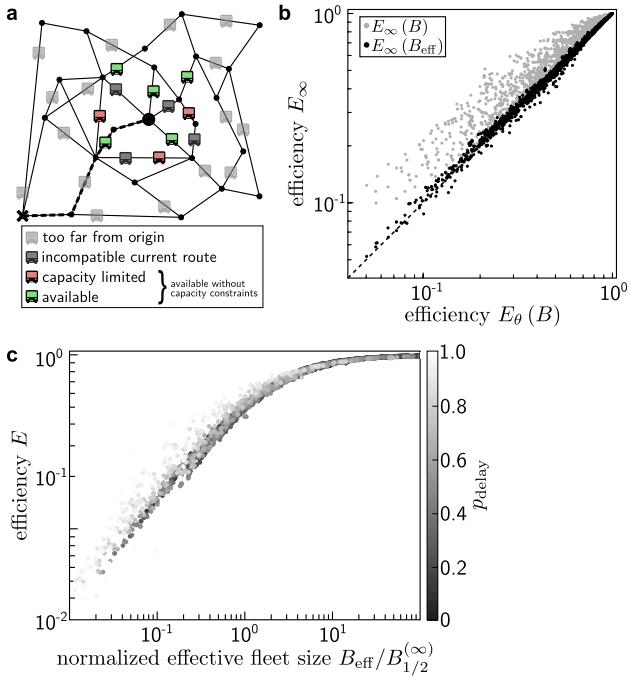

Figure 3. Effective fleet sizes capture the impact of capacity constraints. (**a**) When a new request (black circle, center) arrives, it must be assigned to one of the ride-pooling vehicles in the system. The number of feasible vehicles to serve the request is limited due to the large distance to the origin of many vehicles (light gray) or incompatible planned routes of close-by vehicles (dark gray). In a system without capacity constraints, the request would be assigned to the best of the remaining vehicles. However, a fraction p_{delay} of these vehicles cannot serve the request due to the capacity constraints (light red). This argument suggests that the ride-pooling dynamics of a capacity-constrained system is similar to the dynamics of an unconstrained system with a reduced effective fleet size $B_{\text{eff}} = (1 - p_{\text{delay}})\, B$, Eq. (6). (**b**) The efficiency $E_\theta(B)$ of capacity-constrained systems is approximately equal to the efficiency of unconstrained systems $E_\infty(B_{\text{eff}})$ with the reduced effective fleet size B_{eff} (black dots). Comparing both systems with the same fleet size, the efficiencies differ significantly (light gray). The figure shows results for more than 3000 distinct settings $(\mathcal{T}, B, x, \theta)$ where $p_{\text{delay}} \leq 0.8$. **c** With the normalized effective fleet size as the scaling parameter, the efficiency of capacity-constrained ride-pooling services collapses to the same universal efficiency function as the unconstrained system across different topologies, capacity constraints, and system loads x. Deviations occur when most vehicles are fully occupied, $p_{\text{delay}} \approx 1$ (light dots, see main text). See "Methods" for details on the settings and simulations.

single server[29]. The average queue length $\langle q \rangle$ of this system as well as the full queue length distribution can be computed analytically ([29], see Supplementary Material for detailed calculations).

In the ride-pooling system, the average number $\langle C \rangle = x + 2\langle q \rangle / B$ of scheduled customers per vehicle consists of the number of customers currently transported per vehicle, $\langle O \rangle = x$ since detours are impossible in this setting, and the queues at both nodes, $2\langle q \rangle / B$. The efficiency becomes

$$E = \frac{x}{\langle C \rangle} = \frac{1}{1 + 2\langle q \rangle / (Bx)}, \tag{8}$$

with a similar form as the universal scaling function predicted in[25]. This queueing theoretical prediction (Fig. 4a) becomes exact with $B = 1$ vehicle for sufficiently large load x. For smaller loads, the vehicle becomes idle from time to time as fewer requests enter the system. For larger fleets, $B > 1$, fluctuations of the inter-arrival time lead to slight bunching of the vehicles and less efficient service.

The full queue length distribution from this model also provides direct access to the probability p_{delay} that a request is delayed due to the capacity constraints, i.e. when more than θ requests are waiting at a node when a vehicle arrives (Fig. 4b, see Supplementary Material for detailed calculations). As above, results are exact with $B = 1$ vehicle. For larger fleets, fluctuations of the inter-arrival time and less efficient service result in more delayed requests and slightly larger values of p_{delay} than estimated.

Taking a mean-field approach and assuming that the queueing dynamics and occupancy statistics are identical at every node and vehicles arrive with a constant inter-arrival times in the limit of large fleets, the same approach also provides estimates $p_{\text{delay}}^{(\text{est})}$ for arbitrary networks (Fig. 4b inset). A detailed description of the estimation using a self-consistent solution of approximate queue length and occupancy distributions is given in the Supplementary Material. Differences between the estimated $p_{\text{delay}}^{(\text{est})}$ and the observed p_{delay} occur due to heterogeneities in the networks and the inter-arrival time of vehicles. As an alternative to an equidistant distribution of vehicles and a deterministic inter-arrival time, an exponential inter-arrival time distribution offers a good approximation for

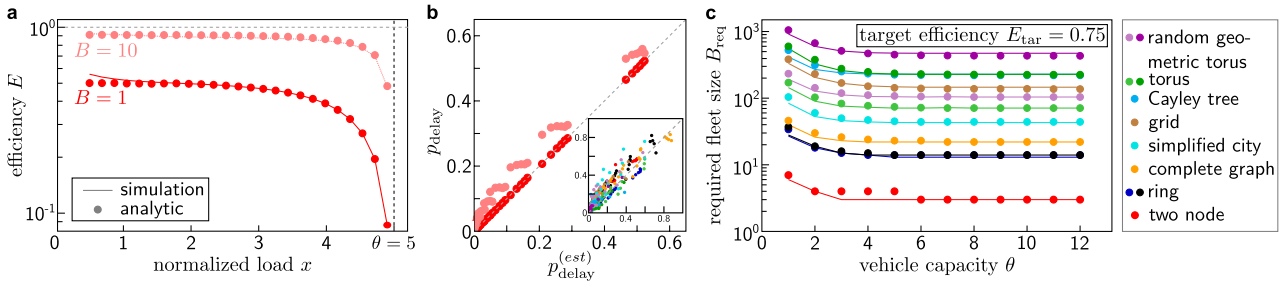

Figure 4. Fleet size prediction for capacity-constrained ride-pooling services. (**a**) Queueing theory predictions (dots) of the ride-pooling efficiency in a minimal two-node setting. The predictions become exact for a single vehicle $B = 1$ (dark red) at high load x where the vehicle is never idle. Small deviations for larger fleet sizes ($B = 10$, light red) reflect the non-equidistant inter-arrival time distribution of vehicles. (**b**) The same queueing theoretical description predicts the scaling parameter $p_{\rm delay}$ for various loads x and capacity constraints θ. (inset) A mean-field approach enables the estimation of $p_{\rm delay}$ in arbitrary networks for large numbers of vehicles (see Supplementary Material for details). (**c**) Prediction (dots) of the required fleet sizes to achieve a desired efficiency $E_{\rm tar} = 0.75$ for various network topologies and capacity constraints compared to direct numerical simulations (lines). These estimates rely only on the universal scaling function $f_{\mathscr{A}}$ and measurements of the scaling parameters $B_{1/2}(\mathscr{T}, x)$ of the capacity unconstrained systems. Colors represent different underlying networks, see "Methods" and Supplementary Material for details on the settings, simulations, and calculations.

the dynamics in large and heterogeneous networks, reflecting the limit of many independent paths along which vehicles may arrive at a node (see Supplementary Material and Supplementary Figure S1).

Together with the scaling function $f_{\mathscr{A}}$, Eq. (5)[25] and the topological factor $B_{1/2}$, this approximation enables us to a-priori estimate the required fleet size to achieve a desired efficiency in a given setting (Fig. 4c). Starting with some fleet size B, we estimate the delay probability $p_{\rm delay}$ and the effective fleet size $B_{\rm eff}$ using the mean-field calculations and compute the resulting efficiency E from the universal scaling function. Comparing this estimate to a desired efficiency $E_{\rm tar}$, we obtain a new estimate for the required fleet size B by assuming the same delay probability $p_{\rm delay}$. Iterating these estimations, the process converges to an estimate $B_{\rm req}$ of the required fleet size to achieve the desired efficiency in the given setting (see Supplementary Material for details). Note that, during this process, the load x changes as the fleet size varies while the vehicle velocity, request rate, and request distribution remain constant. We thus make use of the full range of scaling parameters $B_{1/2}(\mathscr{T}, x)$ of the capacity unconstrained systems to obtain more accurate results. For systems with a high density of requests, the topological factor $B_{1/2}(\mathscr{T}, x)$ may be replaced by the single scaling factor in the limit of large loads $B_{1/2}^{(\infty)}(\mathscr{T})$, which can also be estimated without simulations in many simple networks by counting the number of distinct (shortest) paths[25].

The results of these estimations agree well with the required fleet sizes found from direct simulations in a wide range of network and capacity settings (Fig. 4c). Similar to the analytical calculations above, deviations become larger when $p_{\rm delay}$ is large (e.g. for low-capacity vehicles). However, this usually only occurs for undesirable settings with small target efficiencies or a large number of low-capacity vehicles.

Discussion

The collective dynamics of a ride-pooling fleet determines the potential and actual efficiency of the ride-pooling service[25,27]. Instead of the specific request rate or the normalized loads, we have identified the effective number of available vehicles as the relevant scaling parameter to describe the dynamics of capacity-constrained ride-pooling fleets. This concept of an effective fleet size relates the efficiency of a capacity-constrained ride-pooling system to a system without capacity constraints and recovers the topological universality observed in systems with unlimited capacity[25]. The successful mapping between the collective dynamics of capacity-constrained and unconstrained systems suggests that a similar approach may be able to capture the impact of other constraints limiting the assignment of requests to vehicles, such as heterogeneous request sizes from individual travelers and groups or mixed request types for single (taxi cab) or shared rides.

The universal scaling of the efficiency in systems without capacity constraints is robust across different demand distributions and network topologies (captured in the average trip length $\langle l \rangle$ and the topological scaling factor $B_{1/2}$) as well as for different dispatcher algorithms in the high-efficiency limit[25]. Since our results are based on a direct mapping between capacity-constrained and -unconstrained systems, this robustness directly transfers as well. The mapping between the capacity-constrained and -unconstrained systems only breaks down for large $p_{\rm delay} \approx 1$ when the system is close to overloading, a state that is undesirable regardless of the setting due to long detours or waiting times. Since all arguments and in particular the definition of the ride-pooling efficiency rely on the equilibrium steady state of the ride-pooling dynamics, our results only capture expected dynamics over long times. Changes on timescales faster than the typical service time of a single customer, such as quickly changing or highly correlated demand distributions, strongly varying request rates λ, or quickly varying traffic conditions and vehicle velocities v, cannot be captured in this equilibrium description. Importantly, the scaling of the efficiency captures the dynamics both from the perspective of the provider in terms of the queueing theoretical throughput as well as from the perspective of the customers due to the direct relation to the

average service time (see Eq. (4)). A relevant additional perspective may be the extension of these scaling laws to the reliability of travel times and the distribution of delays beyond the mean-field description considered here. Similarly, while the dimensionless load quantifies when pooling rides becomes necessary, the sustainability of the service in terms of driven distance and emissions is not directly captured in the scaling laws.

The analytic queueing theory model enables the application of this extended universality beyond numerical simulations. While the mean-field calculations for arbitrary networks cannot be expected to be highly accurate in real-life settings that are strongly heterogeneous, our results in principle enable a-priori estimates of required fleet sizes or efficiencies without the need for detailed simulations, complementing existing results[23,25,27,28] and providing a new tool to study the potential of ride-pooling in previously unserviced areas.

Methods

Ride-pooling simulations. We simulate the dynamics of a ride-pooling service with B vehicles traveling with constant velocity v. We set $v = 1$ in all simulations without loss of generality, measuring time in appropriate units. For every vehicle, we store the planned routes as a list of scheduled pick-up and drop-off stops. Over time, vehicles drive along the shortest path between consecutive stops and pick up and drop off all scheduled customers. If a vehicle has no scheduled customers, it becomes idle and does not move until it is assigned a new customer.

Customers place requests to travel from one node i to another node $j \neq i$, distributed uniformly randomly and independently across all nodes in the network. Requests follow a Poisson process in time with an total rate λ across the network.

Each time a new request is made, the dispatching algorithm iterates over all pick-up and drop-off insertions in the planned routes of all vehicles to find the offer that minimizes the arrival time of the request without delaying any previously scheduled customers. In case of multiple options, the secondary and tertiary objectives are the minimization of the time that the customer spends inside the vehicle and choosing the vehicle with the highest current occupancy, respectively. For transporters with limited capacities, only those offers are considered for which the occupancy does not exceed the capacity limit at any time during the trip.

We simulate the dynamics in a variety of different settings described below. Each setting is described by a tuple of fixed parameters including the network topology $\mathcal{G}$, the fleet size B, the normalized load x (or equivalently the request rate λ) and the capacity limit θ that applies to all vehicles.

In every simulation, we first distribute the (initially idle) vehicles uniformly randomly across all nodes of the network. We simulate 2000 B but at least 10^5 requests to obtain an initial equilibrium state. Starting from this state, we enable the measurement of observables and again simulate in steps of 2000 B but at least 10^5 requests. We stop the simulation when the average number of scheduled customers $\langle C \rangle$ over the last 100 time units deviates less than 10% from the total average $\langle C \rangle$ over the whole measurement period. Only for Fig. 2b in the main manuscript, we slowly increase the load by $\Delta x = 0.05$ and simulated for 1000 or 1000 x requests, whichever is larger (1000 x requests correspond to $1000\,\frac{x}{\lambda} = 1000\,\frac{\langle l \rangle}{vB} = 50$ time units with a fleet size of $B = 50$ vehicles and an average requested distance $\langle l \rangle = 2.5$ on the small torus illustrated in the figure).

Model networks. We simulate the ride-pooling dynamics on different street networks $\mathcal{G}$. Nodes of the network correspond to possible pick-up and drop-off locations for customers and edges correspond to streets, with the edge length $l(i, j)$ between nodes i and j denoting the distance between adjacent nodes.

- A *minimal graph* consisting of $N = 2$ nodes with $l(1, 2) = l(2, 1) = 1$.
- A small and a large *ring* with $N = 25$ and $N = 100$ nodes, respectively, where neighboring nodes i and j have the distance $l(i, j) = 1$.
- A *complete graph* with $N = 5$, $l(i, j) = 1$ for all $i \neq j$.
- A non-periodic square lattice (*grid*) with $N = 100$ nodes and $l(i, j) = 1$ for every edge.
- A small and a large periodic square lattice (*torus*) with $N = 25$ and $N = 100$ nodes, respectively, and $l(i, j) = 1$ for every edge.
- A *simplified city* with $N = 16$ nodes, which resembles a spider web. Four rays point outwards from an imaginary center. Four nodes are placed on each ray. On every ray, each node is connected to its neighboring node(s) on the same ray. Furthermore, on each two adjacent rays, the closest nodes to the center are connected to each other, as well as the third-closest nodes to the center. $l(i, j) = 1$ for any two connected nodes i, j.
- A *Cayley tree* with $N = 46$ nodes and $l(i, j) = 1$ for every edge.
- A small and a large *random geometric torus* with $N = 25$ and $N = 100$ nodes, respectively. The networks are generated from the Delaunay triangulation of N points distributed uniformly at random in the unit square with periodic boundary conditions. $l(i, j)$ is given by the Euclidean distance between the connected points i and j with respect to the periodic boundaries.

Measuring p_{delay}. For each request, the dispatcher finds both the best offer O_θ respecting the capacity constraints and the best offer O_∞ ignoring the capacity constraints. We define p_{delay} as the fraction of requests for which the two assignments O_θ and O_∞ differ in terms of the assigned vehicle, the pick-up or the drop-off time. A difference in any of these parameters implies that the best offer in the unconstrained system has become unavailable due to capacity constraints. Note that the probability p_{delay} is a measure over requests for a single vehicle each time, not a direct measure for the fraction of unavailable, fully occupied vehicles.

Data availibility

Data and code underlying the results in the manuscript and the Supplementary Material is availble in the public Github repository 'PhysicsOfMobility/capacity_constrained_pooling'[30], https://doi.org/10.5281/zenodo.6624420.

Received: 16 March 2022; Accepted: 15 June 2022
Published online: 27 June 2022

References

1. Holovatch, Y., Kenna, R. & Thurner, S. Complex systems: physics beyond physics. *Eur. J. Phys.* **38**, 023002. https://doi.org/10.1088/1361-6404/aa5a87 (2017).
2. Barbosa, H. *et al.* Human mobility: models and applications. *Phys. Rep.*. https://doi.org/10.1016/j.physrep.2018.01.001 (2018).
3. Helbing, D., Farkas, I. J. & Vicsek, T. Freezing by heating in a driven mesoscopic system. *Phys. Rev. Lett.* **84**, 1240. https://doi.org/10.1103/PhysRevLett.84.1240 (2000).
4. Erhardt, G. D. *et al.* Do transportation network companies decrease or increase congestion?. *Sci. Adv.* **5**, eaau2670. https://doi.org/10.1126/sciadv.aau2670 (2019).
5. Schröder, M., Storch, D.-M., Marszal, P. & Timme, M. Anomalous supply shortages from dynamic pricing in on-demand mobility. *Nat. Commun.* **11**, 4831. https://doi.org/10.1038/s41467-020-18370-3 (2020).
6. Storch, D.-M., Timme, M. & Schröder, M. Incentive-driven transition to high ride-sharing adoption. *Nat. Commun.* **12**, 3003. https://doi.org/10.1038/s41467-021-23287-6 (2021).
7. Simini, F., González, M. C., Maritan, A. & Barabási, A.-L. A universal model for mobility and migration patterns. *Nature* **484**, 96. https://doi.org/10.1038/nature10856 (2012).
8. Gastner, M. T. & Newman, M. E. J. Optimal design of spatial distribution networks. *Phys. Rev. E* **74**, 016117. https://doi.org/10.1103/PhysRevE.74.016117 (2006).
9. Verma, T., Russmann, F., Araújo, N. A. M., Nagler, J. & Herrmann, H. J. Emergence of core-peripheries in networks. *Nat. Commun.* **7**, 10441. https://doi.org/10.1038/ncomms10441 (2016).
10. Barthélemy, M. & Flammini, A. Modeling urban street patterns. *Phys. Rev. Lett.* **100**, 138702. https://doi.org/10.1103/PhysRevLett.100.138702 (2008).
11. Brelsford, C., Martin, T., Hand, J. & Bettencourt, L. M. Toward cities without slums: topology and the spatial evolution of neighborhoods. *Sci. Adv.* **4**, eaar4644. https://doi.org/10.1126/sciadv.aar4644 (2018).
12. Xu, Y., Olmos, L. E., Abbar, S. & González, M. C. Deconstructing laws of accessibility and facility distribution in cities. *Sci. Adv.* **6**, eabb4112. https://doi.org/10.1126/sciadv.abb4112 (2020).
13. Karamouzas, I., Skinner, B. & Guy, S. J. Universal power law governing pedestrian interactions. *Phys. Rev. Lett.* **113**, 238701. https://doi.org/10.1103/PhysRevLett.113.238701 (2014).
14. Treiber, M. & Kesting, A. *Traffic flow dynamics* (Springer-Verlag, Berlin Heidelberg, 2013).
15. Loder, A., Ambühl, L., Menendez, M. & Axhausen, K. W. Understanding traffic capacity of urban networks. *Sci. Rep.* **9**, 16283. https://doi.org/10.1038/s41598-019-51539-5 (2019).
16. Saberi, M. *et al.* A simple contagion process describes spreading of traffic jams in urban networks. *Nat. Commun.* **11**, 10441. https://doi.org/10.1038/s41467-020-15353-2 (2020).
17. Marszal, P., Timme, M. & Schröder, M. Phase separation induces congestion waves in electric vehicle charging. *Phys. Rev. E* **104**, L042302. https://doi.org/10.1103/PhysRevE.104.L042302 (2021).
18. Dhawan, R., Hensley, R., Padhi, A. & Tschiesner, A. Mobility's second great inflection point. *McKinsey Quarterly*. Accessed 20 June 2022. https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/mobilitys-second-great-inflection-point (2019).
19. United Nations, Department of Economic and Social Affairs. World urbanization prospects: The 2014 revision. Accessed 20 June 2022. https://population.un.org/wup/publications/files/wup2014-report.pdf (2015).
20. United Nations, Department of Economic and Social Affairs. World urbanization prospects: The 2018 revision–key facts. Accessed 20 June 2022. https://population.un.org/wup/Publications/Files/WUP2018-KeyFacts.pdf (2018).
21. McDonnell, M. J. & MacGregor-Fors, I. The ecological future of cities. *Science* **352**, 936. https://doi.org/10.1126/science.aaf3630 (2016).
22. Ramaswami, A., Russell, A. G., Culligan, P. J., Sharma, K. R. & Kumar, E. Meta-principles for developing smart, sustainable, and healthy cities. *Science* **352**, 940. https://doi.org/10.1126/science.aaf7160 (2016).
23. Santi, P. *et al.* Quantifying the benefits of vehicle pooling with shareability networks. *Proc. Natl. Acad. Sci.* **111**, 13290. https://doi.org/10.1073/pnas.1403657111 (2014).
24. 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.* **114**, 462. https://doi.org/10.1073/pnas.1611675114 (2017).
25. Molkenthin, N., Schröder, M. & Timme, M. Scaling laws of collective ride-sharing dynamics. *Phys. Rev. Lett.* **125**, 248302. https://doi.org/10.1103/PhysRevLett.125.248302 (2020).
26. Lotze, C., Marszal, P., Schröder, M. & Timme, M. Dynamic stop pooling for flexible and sustainable ride sharing. *New J. Phys.* **24**, 023034. https://doi.org/10.1088/1367-2630/ac47c9 (2022).
27. Tachet, R. *et al.* Scaling law of urban ride sharing. *Sci. Rep.* **7**, 42868. https://doi.org/10.1038/srep42868 (2017).
28. Vazifeh, M. M., Santi, P., Resta, G., Strogatz, S. H. & Ratti, C. Addressing the minimum fleet problem in on-demand urban mobility. *Nature* **557**, 534. https://doi.org/10.1038/s41586-018-0095-1 (2018).
29. Bailey, N. T. J. On queueing processes with bulk service. *J. R. Stat. Soc. B* **16**, 80. https://doi.org/10.1111/j.2517-6161.1954.tb00149.x (1954).
30. Zech, R. M., Molkenthin, N., Timme, M. & Schröder, M. *Code and data accompanying Collective dynamics of capacity-constrained ride-pooling fleets*. https://doi.org/10.5281/zenodo.6624420 (2022).

Acknowledgements

We thank Debsankha Manik and Philip Marszal for helpful discussion and suggestions. M.T. acknowledges support from the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) through the Center for Advancing Electronics Dresden (cfaed). M.S. and M.T. acknowledge support from the Volkswagen Foundation (VWF, Volkswagenstiftung) under Grant. No. 99 720.

Author contributions

M.S. and N.M. conceived the research, R.M.Z. performed the simulations and analytical calculations and created the figures supported by M.S., all authors analyzed the results and wrote the manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-14960-x.

Correspondence and requests for materials should be addressed to M.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.